國立臺灣大學理學院地理環境資源學系

博士論文

Department of Geography
College of Science
National Taiwan University

Doctoral Dissertation

多點群聚現象的跨尺度特性

The scaling properties of point clustering phenomena

陳威全

Wei Chien Benny Chin

指導教授：溫在弘博士

Advisor: Tzai-Hung Wen, Ph.D.

中華民國 107 年 7 月

July, 2018

doi:10.6342/NTU201802404

doi:10.6342/NTU201802404

國立臺灣大學博士學位論文
口試委員會審定書

多點群聚現象的跨尺度特性
The scaling properties of point clustering phenomena

本論文係陳威全君 (D03228002) 在國立臺灣大學地理環境資源學系完成之博士學位論文，於民國 107 年 7 月 11 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

所　長：

doi:10.6342/NTU201802404

doi:10.6342/NTU201802404

誌謝

轉眼間，來臺已近十一年。台北市也就這樣成爲我這輩子至今居住最久的一個城市。從 2007 年 9 月第一次踏入地理系館的門開始，在這個系館也生活了這十一年。這段日子中，從不適應到很混，然後再從後悔自己很混而開始認眞，然後也忘了是什麼機緣下開始念研究所，到現在從寫程式轉到寫論文的這一段日子中，陸陸續續，大大小小的各種事務，都受到地理系的大家的協助。以下無法一一列舉，所以在這先感謝各位。

感謝指導教授溫在弘老師，從大三的計量地理開始，到後來碩士班、博士班一路以來的培養，讓我從不懂寫程式到現在什麼都要用程式來處理，包括這本論文的文本也是用程式來寫出來，以及引導我寫文章，甚至教會了我很多學術生涯上的人生哲理。認眞的說，沒有溫老師就不會有現在的我！感謝我的學術指導委員，系上的林楨家老師以及長庚大學的黃崇源老師，在研究上，包括跟我論文有關的或是跟我其他研究有關的方面都提供了很多的支持、意見以及幫助。感謝政大的余清祥老師、海大蔡宇軒老師、還有遠從丹麥 Aarhus 來的 Clive 老師，百忙之中幫我評閱我那又長又用菜菜的英文寫的論文，並且在不知道會不會放颱風假的那一天來參與我的口試；也感謝瑪莉亞颱風跑的比較快，不影響我口試日期。感謝大四以及更之前的那幾年指導我的李美慧老師，李老師帶著我從對研究完全沒有概念，到開始觀察渦蟲，然後完成科技部大專生計畫以及大四專題，讓我對後來的人生規劃有很大的轉折。雖然我後來的方向轉移了很多，不過那時候如果沒有李老師的引導，也許我還是當年那個很混的樣子吧！當然還要感謝很多系上的各位老師們，無論是在課堂上、演講上，或是各種茶餘飯後的時候的分享與閒聊，都對我在地理這個大領域上的了解有很大的幫助。也要感謝系上的同學們與學長姐學弟妹們，尤其是 501 室的一起打拼的大家，無論是研究上還是生活上都給了我很大的幫助。B96 的大家一起開考試前的讀書會、一起出野外實查；R00 的大家一起撐過寫論文的日子；D03... 好像也沒什麼回憶哈；還有我碩班畢業後開始沒辦法用學號來記的大家一起被釘的很慘的朋友們。大家好好保重！

最後，要感謝支持我在高中畢業後跑得這麼遠來唸書，還一念就是十一年的家人，謝謝你們的體諒與包容。最後最後，壓軸要感謝的，是我的愛妻素華，謝謝妳願意相信我、陪著我、支持我把這博士班唸完。

doi:10.6342/NTU201802404

doi:10.6342/NTU201802404

摘要

傳統針對點形態分析的空間統計方法主要聚焦於其全域或區域形態的探討，而忽略了點資料的整體分佈趨勢，這整體分佈可透過一個跨尺度過程中被觀察與描述。過去空間統計研究提出，由多點所組成的點分佈則可能呈現出會隨着尺度而改變的空間形態，亦即可調整空間單元問題 (MAUP) 。另一方面，碎形分析研究指出，點形態在跨尺度過程的初期會以固定的速率發生變化，而經過某一尺度後此速率會降低。發生速率轉換的尺度 (即臨界尺度) 是可反映點的整體分佈趨勢 (即巨觀形態) 的最大尺度，亦即從全域形態開始逐漸提高尺度的過程中，其形態因 MAUP 而發生變化的速率是固定的，而在臨界尺度以上的尺度遞增則只對一些區域造成形態上的影響。

本研究提出一個跨尺度分析架構以偵測臨界尺度，並建立一套將原始點分佈加總成巨觀形態的流程架構。本研究進行了三項實驗，包括兩個針對理論分佈的實驗以及一組實證資料的分析。這三項實驗分別以：(1) 探討跨尺度過程中的加總效果；(2) 探討單核心群聚特性對跨尺度分析的影響；以及 (3) 探討眞實資料中所反映的巨觀形態與原始分佈之間的差異。結果反映在臨界尺度下的整併點分佈已足夠反映原始點資料的空間形態；群聚的大小與組成群聚的點數量對於跨尺度分析結果呈對數變化的關係；整體分佈趨勢可悲臨界尺度下加總點所形成的巨觀形態所捕捉。

在過去常見於瞭解點分佈形態的全域或區域空間統計研究中之外，本研究不但提供了一個新的分析工具，也提供了一種對於檢視點分佈形態上的新視角。這個分析架構，包括臨界尺度的偵測，以及巨觀形態的呈現方式，對於空間點資料的資料探索以及地圖視覺化將有所幫助。

關鍵字：跨尺度、點分佈、點區塊四分樹、碎形視角、群聚現象、巨觀形態

doi:10.6342/NTU201802404

doi:10.6342/NTU201802404

Abstract

Conventional spatial statistics analysis reveal either the global or local spatial pattern for point distribution but ignore the big picture of the point data, which can be observed through a scaling process. As discussed in spatial statistical studies, point distribution may show a scale-dependent spatial pattern, namely the modifiable areal unit problem (MAUP). On the other hand, as suggested in fractal analysis studies, the point pattern change at a consistent rate at the beginning of the scaling process, and the rate decrease after a scale. This scale of rate shifting (namely critical scale, CS) is the finest scale that can capture the big picture of the point distribution (namely macro pattern), i.e. the changes of pattern due to MAUP have the same effect since the global level, and the higher scales after CS will only affect the pattern of partial area.

In this study, a scaling analysis framework was proposed to identify the critical scale, and a set of aggregation procedure was designed to convert the original point dataset into the macro pattern. Three experiments were conducted to test the scaling analysis framework, including two experiments on theoretical distributions and one on empirical point distributions. The three experiments were designed: (1) to test the aggregation effect on scaling process; (2) to test the influences of mono-centric clustering properties to the scaling analysis; and (3) to illustrates the contrast between macro patterns and the original distribution of empirical data. The results suggested that the aggregation point on critical scale could capture most of the spatial properties from original data; the area and number of point formed a logarithm relationship

doi:10.6342/NTU201802404

with the critical and final scale; and the big picture of the point distribution could be captured by the macro pattern of the aggregated points on critical scale.

Aside from the conventional understanding of point pattern as discussed in the previous global or local spatial statistics methods, this study provides not only a new tool but also a novel perspective of viewing the point distribution. This analysis framework, including the critical scale identification and macro pattern aggregation, can be useful for spatial point data exploration and map visualization.

Keywords: Scaling, point distribution, Point-Region Quadtree, fractal perspective, clustering phenomenon, macro pattern

doi:10.6342/NTU201802404

Table of contents

doi:10.6342/NTU201802404

doi:10.6342/NTU201802404

doi:10.6342/NTU201802404

List of Figures

doi:10.6342/NTU201802404

doi:10.6342/NTU201802404

List of Tables

doi:10.6342/NTU201802404

List of Algorithms

List of Equations

doi:10.6342/NTU201802404

doi:10.6342/NTU201802404

Chapter 1

Introduction

1.1 Motivation

Unlike other types of spatial data, the point itself does not imply any scale information (Goodchild and Mark, 1987; Cressie, 1993; Upton et al., 1985), i.e. scale-invariant. The clustering phenomenon in points distribution can be observed in different scales for the same set of point data (Goodchild, 2011). Thus, a cluster in a higher scale can be found within a cluster at a lower scale. For example, the points in both sub-figures of Figure 1.1 shows the cases distribution of a disease in a city. Figure 1.1a has a larger region ($20km \times 20km$). The distribution shows a highly clustered zones bottom-middle area; the area above the highly clustered zone has a moderate clustering pattern; the top-right area has only a small number of points scattering dispersively. Figure 1.1b has only a quarter of the previous region ($5km \times 5km$), which is zooming into the clustered zones of the Figure 1.1a. The distribution also includes a highly clustered zones at the right part of the figure, a moderate clustering zone at the top part, and fewer points appear at the right side of the figure. The patterns of highly clustered, moderate clustered, and dispersion were observed in both of the distributions, indicating that the self-similarity of clustering pattern exists in the distribution of geographical entities (Raines, 2008; Agterberg, 2013). This also indicates that points clustering pattern can be a scale-invariant phenomenon(Goodchild and Mark, 1987), rather than always to be a scale-dependent phenomenon, i.e. the fixed pattern that can be determined under only a certain scale or a pat-

doi:10.6342/NTU201802404

tern that should be observed under a specific scale (Agterberg, 2013; Frankhauser, 2015).

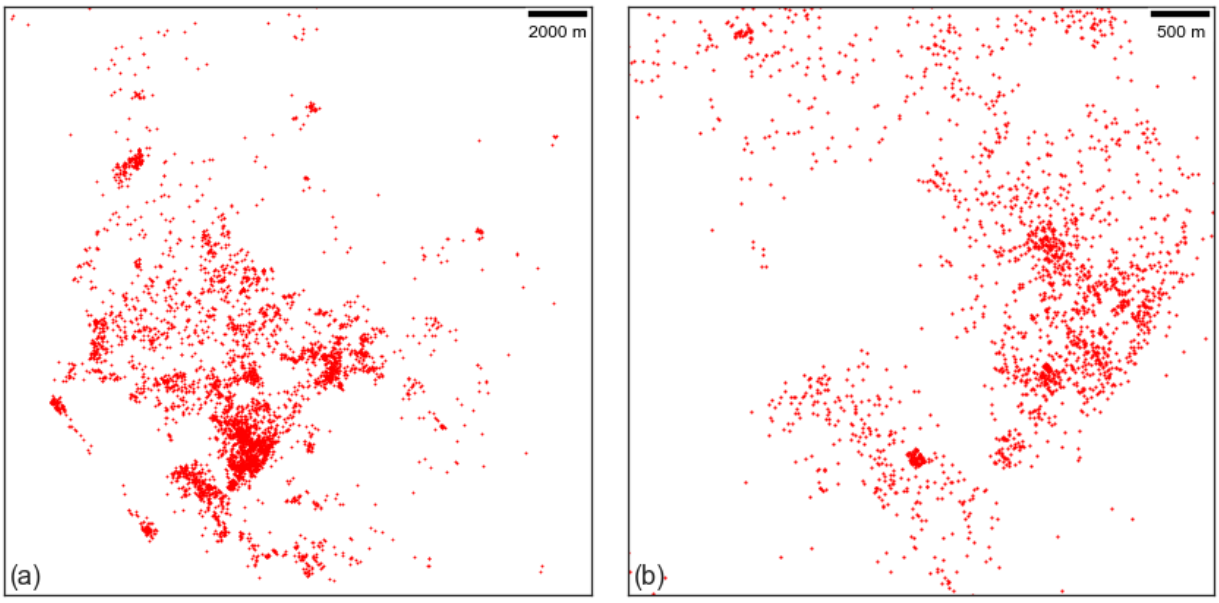


Figure 1.1: Clustering phenomenon of points distribution (Dengue Fever in Kaohsiung City, Taiwan) in different scale: (a) $20km \times 20km$ area, and (b) $5km \times 5km$ area.

Spatial heterogeneity is one of the geographical phenomena that often appears in the spatial distribution of geographical entities (Gatrell et al., 1996), and which is also the main cause of the scale-invariant clustering phenomenon. In the discussions about point distribution, the spatial heterogeneity is focused on the distribution that points are concentrated at some small parts of the study area, leaving some other parts empty, which is also known as the clustering patterns (Kulldorff and Nagarwalla, 1995). Most of the previous studies relied on the analysis of the distance between points (Fotheringham and Zhan, 1996; Diggle, 2013), and tried to identify the cluster areas by using circles, ellipses, or zones from a surface (Nakaya and Yano, 2010; Anderson, 2009; Gerber, 2014), or joining the points into grids or administrative regions (polygons) for the analyses using aggregated approaches (Anselin, 1995; Ord and Getis, 1995). The approach behind these studies mainly focused on the discussion of a fixed scale (e.g. a fixed searching distance, a fixed number of points to be announced as a cluster, or fixed level of administrative regions).

Previous approaches that were designed in order to search for clusters in a specific scale, i.e. scale-dependent, defined a cluster as a determined pattern by using a Euclidean

doi:10.6342/NTU201802404

distance (or any cost measurements, e.g. Manhattan distance, trip length, traveling time) as the major parameter (Kulldorff, 1997; Lee et al., 2014). This can be useful in practical applications. But, this ignored the nature of points distribution that the clustering phenomenon may be scale-invariant (Agterberg, 2013; Frankhauser, 2015). In other words, the clustering of geographical points can be a cross-scales phenomenon. Analyzing and visualizing the clusters of points under certain scales may result in a biased understanding because **the pattern revealed only a partial truth about the spatial patterns**.

The analysis based on a specific scale reveals the partial truth of the spatial patterns, and which may not be true for other scales. This is the scaling issue of the popular **modifiable areal unit problem (MAUP)** (Openshaw, 1983; Fotheringham and Wong, 1991), which means that the statistical analysis based on different aggregation of points can lead to a different conclusion. Or in other words, the observed patterns of point data is scale-dependent. In geographical pattern analysis, scaling issue of MAUP indicates that the spatial pattern may change if the grouping level is different, e.g. between county level (larger groups) and basic statistical unit level (smaller groups). While MAUP is a crucial problem in analyzing spatial patterns, the question about how does the pattern changes over scales and to which level of scale the changes stop, or if these scales exist, become key issues for understanding the spatial pattern.

Scaling from the lowest scale (worst resolution) to the highest scale (finest resolution), the point pattern will change and the become clearer with the increasing of scale. That is, more spatial information can be derived from the point pattern with higher resolution. For example, the demonstration in Figure 1.2 showed the points count in each cell on different scales. The information of spatial patterns increased, in terms of showing where there are more point events, with the increasing resolution at the first several scales. But, as the resolution continues to be increased, the distribution of the occupied cells are becoming more like the points distribution, which is useless for describing and measuring the spatial patterns. In other words, the increment rates of the spatial information slow down after a turning point, i.e. a critical scale. This situation is discussed as the roll-off effect in fractal studies on point data (Raines, 2008; Agterberg, 2013).

doi:10.6342/NTU201802404

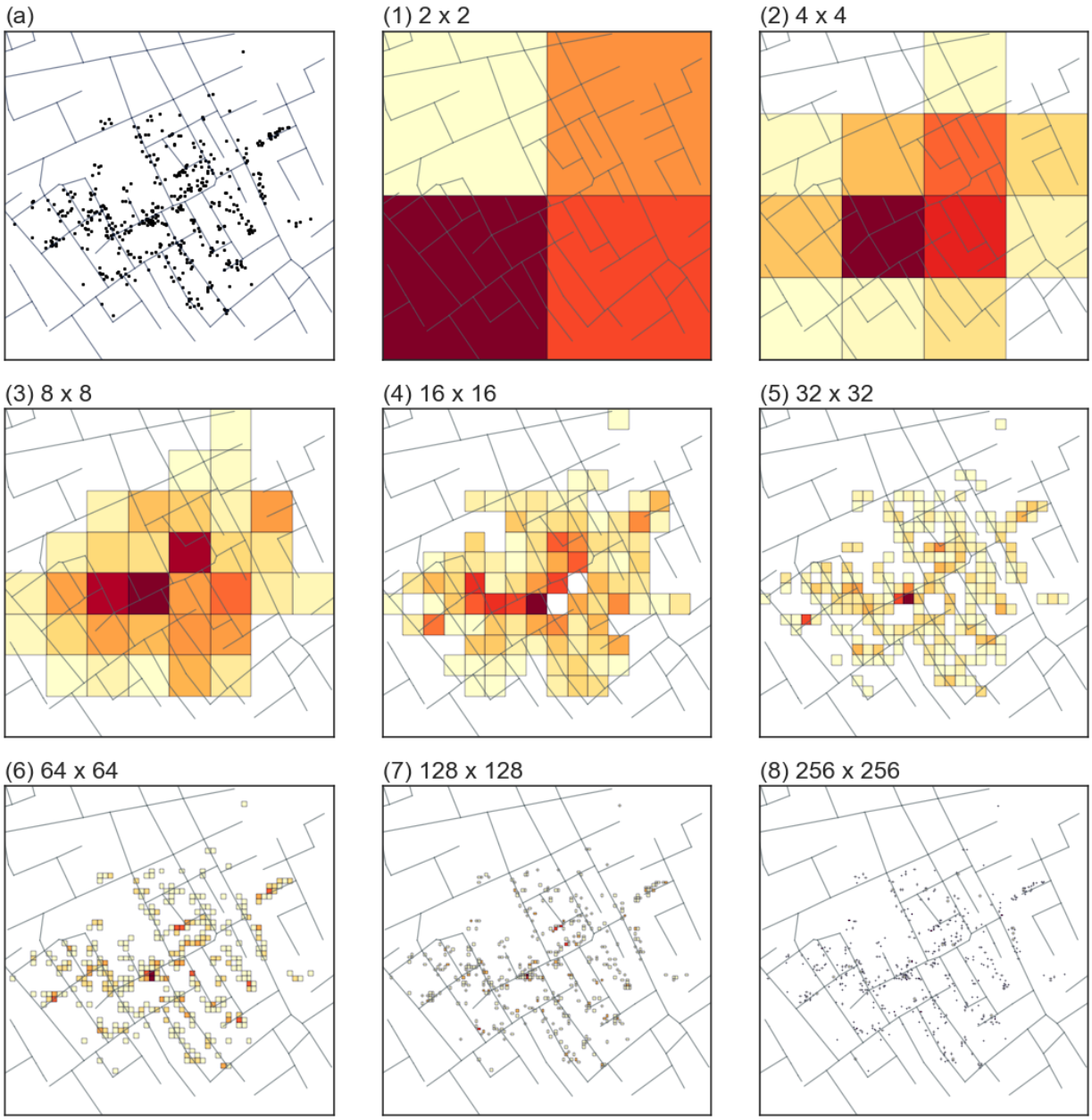


Figure 1.2: A demonstration of points scaling using John Snow cholera data (Arribas-Bel et al., 2017). The colors indicate the number of cases within each box, the deeper the color, the more cases the box contains.

Previous studies about the **fractal dimension in geographical features** provided another approach for analyzing the scaling properties of spatial distribution (Mandelbrot, 1967; Frankhauser, 2015). The key concept of geographical fractal analysis is to reveal the relationships between spatial scales on the distribution patterns (Mandelbrot, 1967; Agterberg, 2013; Goodchild and Mark, 1987). The fractional aspect of geographical distribution has been an important perspective for discussing geographical phenomena since

doi:10.6342/NTU201802404

the last two decades (Goodchild and Mark, 1987; Batty et al., 1989; Batty and Longley, 1994; Keersmaecker et al., 2003; Batty, 2007). Previous geographical studies mainly used fractal analysis approach to analyze urban sprawl: Batty et al. (1989) modeled the urban growth and form as a tree-like dynamic development process; Yu and Ng (2007) and Sutton (2003) analyzed satellite images of different time periods to measure the dynamics of urban sprawl; Terzi and Kaya (2011) analyzed urban sprawl using demographic data. The fractal analysis of point distribution was also discussed comprehensively in geophysical studies. Carlson (1991), Blenkinsop and Sanderson (1999), Ford and Blenkinsop (2008), and Gumiel et al. (2010) used fractal methods to understand the locations and spatial attributes of the mineralization; Hayakawa et al. (1999), Kagan and Jackson (1991), and Varnes and Bufe (1996) analyzed the fractal features of the distribution of seismic activities.

The conventional spatial statistics methods for point data analysis includes two major types, i.e. the global and local aspects. The global analysis is used to treat all point as equal and to derive some indexes to describe the point pattern, e.g. clustered or not clustered. The local analysis is used to investigate the neighboring points and to present the local spatial variations, e.g. the locations of hot spots and cold spots. Both of these treats all points as distinct and equal entities that may or may not be related to each other depends on their distances. The global analysis may be dominated by some partial locations that have a larger set of points, which have stronger relationships because all of them were equally weighted in the calculation; and ignore the spatial patterns of other areas with a smaller set of points. On the other hand, the local analysis concentrates distinctly on the local neighborhood areas while calculating for one location, and ignore the spatial pattern of the other areas. Therefore, both the global and local analysis cannot be used to capture the big picture of the point distribution.

In order to represent the big picture of point distribution, i.e. treating all points as a whole distribution object rather than a combination of distinct point entity, this study aimed to investigate the scaling process, to propose an analysis framework to reveals the scaling properties including the critical scale, and to reconstruct the big picture based on

doi:10.6342/NTU201802404

the scaling properties. The scaling properties between scales can be analyzed utilizing the fractal analysis perspective (Sémécurbe et al., 2016; Agterberg, 2013). Previous fractal studies focused mainly on the fractal geometries, i.e. shapes and distributions. From the perspective of clustering phenomena (Upton et al., 1985; Cressie, 1993), point intensity is one of the main properties of clusters (Nakaya and Yano, 2010; Chainey et al., 2002; McCord and Ratcliffe, 2009), which was neglected by most of the spatial fractal analysis that focused mainly on the shapes and size of area aspects (Batty et al., 1989; Carlson, 1991). The intensity aspect can also be formulated as another fractal dimension (Raines, 2008; Agterberg, 2013), indicating that it is also experiencing the scaling properties. Based on the fractional perspective, this study focused on the scaling properties of the point distribution clustering phenomenon.

1.2 Background and related studies

1.2.1 Modifiable areal unit problems

While points do not imply scale information, they can be analyzed using different types of scales. The aggregation of points to polygons of different scales will encounter an issue called modifiable areal unit problems, which is often referred to as MAUP (Openshaw, 1983). MAUP is a situation that while we observe the distribution of the same data in two different scales or different zoning systems, the resulting patterns can be completely different. The two aforementioned situation, i.e. in different scales and in different zoning systems, are forming the scaling effect and zoning effect of MAUP, respectively. Figure 1.3 shows a demonstration of scaling effect. In Figure 1.3b, the points distribution seems like concentrated at the bottom left and bottom area; in Figure 1.3c, it shows that the top right cell has the largest number of points. The scaling effect means that while we merge several small polygons into a larger polygon, the scales become smaller, and the distribution of the patterns changed (Fotheringham and Wong, 1991).

Previous studies stated that cross-scales inferences caused problems on the understanding of the distribution (Goodchild, 2011). These problems that were caused by scaling ef-

doi:10.6342/NTU201802404

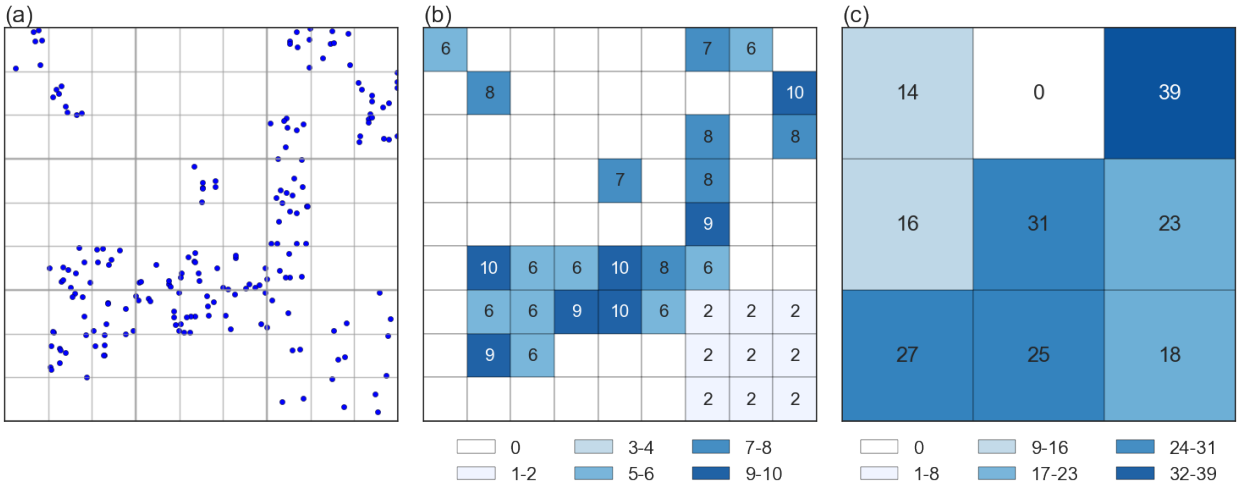


Figure 1.3: A demonstration of the scaling effect of MAUP: (a) the locations of points, (b) grouping the points into a larger scales (smaller boxes), and (c) grouping the points into a smaller scales by aggregating the smaller boxes into larger boxes.

fect is similar to another problem called ecological fallacy (Openshaw, 1984), which was inferences of the results from the observation of individual behaviors to the population behaviors, and vice versa (Openshaw, 1984; King, 1997; Goodchild, 2011). Therefore, it is recommended to analyze a phenomenon on an appropriate scale for the corresponding issue and to keep the discussions in the same level of scale (Taylor et al., 2003). The studies of cross-scales (Kulldorff, 1997; Sémécurbe et al., 2016; Tannier and Pumain, 2005; Thomas et al., 2008) and integrated scales through downscaling or upscaling (Luoto and Hjort, 2008; Zhu et al., 2001; Zhang et al., 1999; Goodchild et al., 1993) can still be done through a comprehensive study design.

1.2.2 Fractal analysis

To reveal the fractal patterns from an empirical data-set, the box-counting approach is the most intuitive method that can be adopted for different types of spatial data (Carlson, 1991; Agterberg, 2013; Frankhauser, 2015), including points and polygons. In simple words, the concept of box-counting is to continue splitting the areas into boxes, and then smaller boxes, iteratively. If a fractal pattern or scaling factor exists in the data, i.e. self-similar patterns appear from scale to scale, the scaling factor can be observed through the iteratively splitting (scaling) process. The observation can be converted into a quantitative

doi:10.6342/NTU201802404

index of fractal dimension by fitting the slope of the box-size to box-count figure (e.g. Figure 1.4, see Section 2.4 for more details) (Raines, 2008; Tannier and Pumain, 2005).

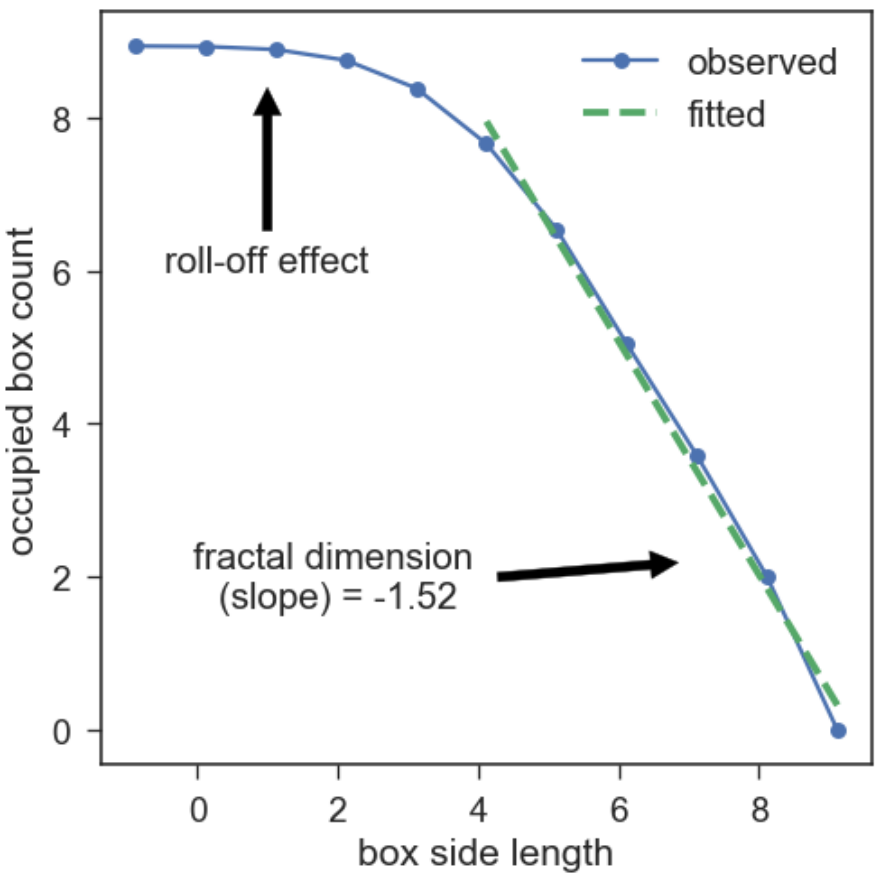


Figure 1.4: A demonstration of the calculation of fractal dimension and the roll-off effect. This demonstration used the John Snow cholera point data as an example. According to the box-counting method, the fractal dimension can be calculated as the slope of the fitted straight line on the logarithmic occupied box against the box side length. On the other hand, part of the occupied box with a smaller box side length cannot be fitted on the straight line, which indicates the occurrence of the roll-off effect.

Previous studies on revealing the scaling factor in point distribution show a common finding that roll-off effect does exist while the scale goes beyond the higher (finer) resolution limit (Agterberg, 2013; Raines, 2008). In simple words, the roll-off effect indicates that starting from the scale level, the pattern does not follow the self-similar trend (e.g. Figure 1.4). Therefore, the roll-off effect would dilute and deviate the observation for the scaling properties by including shifted counting values.

From the perspective of spatial analysis, the scaling factor reveals the macro scaling pattern of the whole distribution, that is a global view of the pattern. On the other hand,

doi:10.6342/NTU201802404

the scales that experienced the roll-off effect indicated that the finer scales are needed for some areas that contain micro distribution pattern, and which pattern exceed the fractal pattern of the global distribution. This situation was observed in previous studies and may be explained as the secondary fractal pattern, i.e. the second part of the bi-fractal pattern (Agterberg, 2013; Raines, 2008). In empirical studies of geographical features, such as urban forms or population distribution, a bi-fractal pattern is often observed (White and Engelen, 1993; White et al., 2015).

1.2.3 Point-region quadtree

While the box-counting approach iteratively splitting the areas into smaller boxes, this process is similar to the mechanism of the Point-Region Quadtree (PR-Qtree) data structure. PR-Qtree is a quadtree data structure for storing and indexing the two dimensions data, and which is often used for the spatial querying of the geographical point data (Orenstein, 1982; Samet, 1984). In simple words, the mechanism of PR-Qtree keeps splitting the boxes with more than one point, until each box contains at most one point (see Section 2.1). Therefore, this can be used for implementing the box-counting approach with some modification (see Section 2.2). In addition, the depth structure of the PR-Qtree indicates the scaling levels. Thus analyzing the PR-Qtree structure can also be used to explain the scaling factor of the point spatial distribution.

1.3 Aims

To sum up, the box-counting approach was used in previous studies for uncovering the scaling factors in point distribution, which analysis process could be captured and simplified through the PR-Qtree mechanism. Therefore, this study intended to adopt the PR-Qtree mechanism for establishing a framework of study for analyzing and uncovering the scaling properties.

The roll-off effect and bi-fractal pattern were frequently observed from empirical spatial point data. Thus, a degree of scale exists between the **global scaling process** and **local**

doi:10.6342/NTU201802404

scaling process. Through an additional calculation based on the box-counting approach, this study proposed an analysis framework which will test the fitness of these scaling phenomena and identify the turning point of scale that is in between the global and local scaling phenomena, namely **critical scale**.

In the perspective of the fractal pattern, the critical scale indicates the scales that the self-similarity process through scaling from the global scale (i.e. the distribution as a whole) towards a finer resolution, has stopped and turned into the secondary fractal pattern. In other words, viewing the distribution by scaling from a lower scale (coarse resolution) to a higher scale (fine resolution), the pattern changed with the scaling, but the amount of changes follows a fixed value of speed, and the changes in geometry follow a fixed pattern, i.e. the self-similarity. The scaling starting from the critical scale, on the other hand, the speed of changes will become a lot slower, and the big picture of the distribution pattern will be maintained, i.e. the changes of the spatial pattern will not be dramatical. Therefore, this situation is connected with the scaling issue of MAUP in the perspective of the geographical pattern.

From the point of view of the geographical pattern, scaling problem of MAUP indicates that the spatial pattern may change if the grouping level is different. While the critical scale indicates that the pattern revealed from the higher scale than the critical scale will not have great changes, it means the sizes smaller than the critical scale will not change the finding of spatial pattern. On the other hand, following the discussion on the fractal perspective, the spatial pattern will change with a fixed rate as the scale is becoming smaller (e.g. larger spatial units). This situation makes that the spatial units at the critical scale are the finest that will keep the spatial patterns as the original point distribution. In other words, if we aggregate the points on the critical scale, the pattern should be similar with the patterns of the aggregation points with a higher scale, and this should include the highest scale, which is the point distribution itself.

The objective of this study is therefore to solve the question: ”*to what scale a point distribution is the optimal spatial level for capturing the macro pattern with the most detail information*”. Then, this study intended to further aggregate the point distribution into a

doi:10.6342/NTU201802404

smaller set of points that capture the most point pattern from the original point distribution. The specific aims of this study are listed as below:

1. To establish an analyzing framework to search for the critical scale;
2. To test the capability of capturing the original spatial pattern using the aggregated points of the critical scale;
3. To analyze the changes in critical scale in response to the changes in clustering properties with a single cluster;
4. To investigate the macro pattern and micro pattern with the empirical dataset.

While the aim of this study is to establish an analyzing framework for identifying the critical scale, this dissertation is organized as Figure 1.5. First, the motivation, background, and aims of the study were discussed in the first chapter. Second, The proposed framework was described in the second chapter. Then, three experiments were conducted to test the framework, followed by the discussions and conclusion chapters.

To test the aforementioned framework and to evaluate the capability of the identified critical scale in terms of capturing the spatial pattern of the original distribution, an experiment using three theoretical distribution (disperse, clustered, and random) were performed, and the aggregated points extracted based on each scale were tested with several basic spatial pattern analysis method of point data, and the results were compared with the findings from original distribution.

The second experiment was designed to test the influences of the clustering phenomenon in the identification of scaling properties. In this experiment, a synthetic single circle shape clustering model was used to test the three aspects of clustering phenomenon, which included the location of clusters, the spatial size of clusters, and the density of clusters.

While one of the aims for the third experiment was to applied the proposed framework on the point distribution in the real world, the major objective was to differentiate the macro pattern and micro pattern from the original dataset and to demonstrate the differences of the macro pattern and the original spatial pattern. Three sets of point data were

doi:10.6342/NTU201802404

used as the case studies, including the location of post offices, photocopy shops, and beverage shops, in a study area within the high population density areas of Taipei City. These three datasets represented different types of distribution: the post offices are dispersed due to the needs of maximizing coverage; the photocopy shops are clustered mainly because of the huge demand for photocopying and document printing is near to the clustered universities in Taipei City; the beverage shops shows a possible random distribution in the densely populated study area. These datasets are complex in comparison to the synthetic point distribution as simulated in the previous two experiments; the underlying point distributions are known and clear.

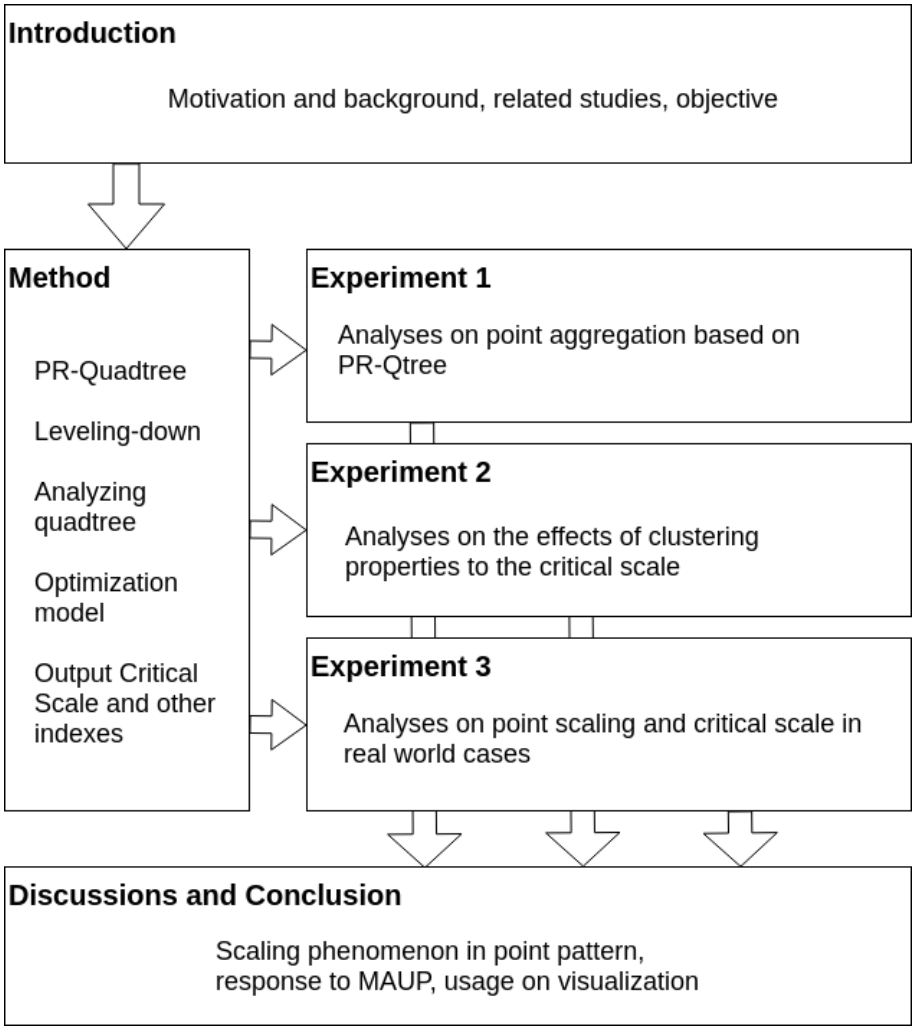


Figure 1.5: The organization structure of dissertation.

doi:10.6342/NTU201802404

Chapter 2

The point scaling analysis framework

Key concepts

In this study, a point scaling analysis framework is proposed to analyze the point scaling process and determine the critical scale. key concepts are related to the development of this framework, which was described and defined in the following:

The scaling phenomenon of points is the process of observing the point distribution from a global view to local view. In this process, the distribution can be described based on different level of scale. Thus, the trends of these observations through the scales can be shown and discussed.

The macro pattern is the observation focusing on the entire study area and treating or viewing all of the points as a whole object, i.e. the global distribution. This is slightly different from the conventional working definition of the global analysis in spatial studies, i.e. as the opposite concept of local analysis, which focused (also) on the global distribution and attempted to derive a conclusion about the spatial distribution, e.g., if the distribution is a significant clustering pattern. The popular global analysis methods include the nearest neighbor analysis, K-function, and Moran's I. While the *macro pattern* in this study indicates a whole view of the pattern, it does not mean to measure the distribution as an index or to determine if the distribution presents a random or clustered pattern; it focuses on the description of the distribution of points altogether, or the relatively micro patterns that are within the macro

doi:10.6342/NTU201802404

distribution.

The micro pattern is the description of distribution that focuses on the patterns that occurred locally and partially, which does not affect the understanding of the global view. The focus of micro patterns is paid to those distributions that do not have the ability to impact the macro pattern. In other words, the micro patterns exist when there are some points are too close to each other, and which situation does not always happen within the point distribution of the specific study. Therefore, the concept of micro pattern is also different from the working definition of the conventional local spatial analysis, which aims to describe the local distribution for each of the spatial features. The terms micro patterns in this study are describing the observation by viewing the point scaling process on the higher scales.

The critical scale The critical scale is a turning point within the process of point scaling, on which scale the micro patterns can be merged and aggregated, to present the macro pattern. The critical scale is defined as the scale that can differentiate the macro pattern from the whole distribution.

Framework design

In this study, an analysis framework for analyzing the scaling pattern of point distribution and to identify the critical scale is designed and introduced. This analysis framework uses point-region quadtree (PR-Qtree) as the foundation. Each of the depth in a PR-Qtree indicates a specific size of resolution cell; in other words, the depth of PR-Qtree represents a spatial scale. Therefore, PR-Qtree is suitable and convenient to be used for the analysis of scaling properties.

PR-Qtree was formally designed for optimizing the point querying process in two-dimensional space. To use it as the foundation for the box-counting method, some procedure is needed to prepare the tree for this specific calculation. Box-counting method is used to calculate the fractal dimension for point data (normally two dimensions). Since the secondary fractal dimension exists in non-regular point distribution (including clustered and random points), this study designed a series of procedure to automatically identify

doi:10.6342/NTU201802404

the turning point, which represents the critical scale that is in between the primary fractal dimension and secondary fractal dimension.

The framework of the calculation procedures are shown in Figure 2.1. The five procedures are discussed in the following sections. The basic **PR-Qtree** is described in the following Section 2.1. One additional modification of the PR-Qtree, namely the **leveling-down process**, is discussed in Section 2.2; the basic **box-counting method** is described in Section 2.3; one optimization process that is based on the box-counting method and an additional index to search for the **critical scale** is discussed in Section 2.4; finally, a set of **measurements** for capturing the scaling properties including the critical scale is described in Section 2.5.

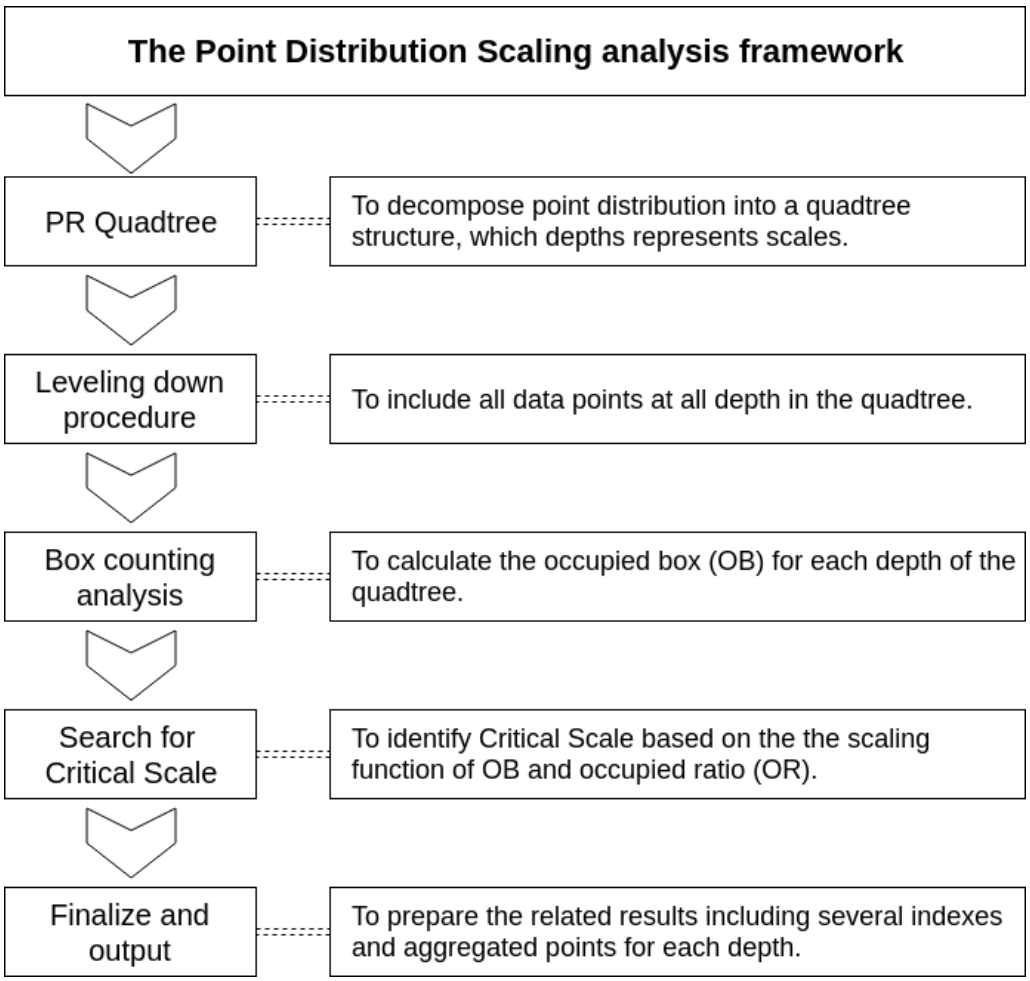


Figure 2.1: The scaling analysis framework.

doi:10.6342/NTU201802404

2.1 Point-Region Quadtree

Point-Region Quadtree (Orenstein, 1982; Samet, 1984), which is one of the **Quadtree data structures**, is known as a spatial indexing data structure that increase the performance of point location queries and range queries (Figure 2.2). Similar to other tree-like approaches (Venables and Ripley, 2002), PR-Qtree divided the spatial area into different parts to store and for further analyzing the data. In simple words, PR-Qtree divided a two dimensions study space into equal size hierarchical quadrants in order to store each of the points in one of the quadrant cell. For the purpose of branch labeling (and the orders of the branches), the id of each quadrant is labeled started from the north-east quadrant as 1, and goes counter-clockwise until the south-east quadrant as 4 (Figure 2.3a); the position of each quadrant in the tree structure is shown in Figure 2.3b.

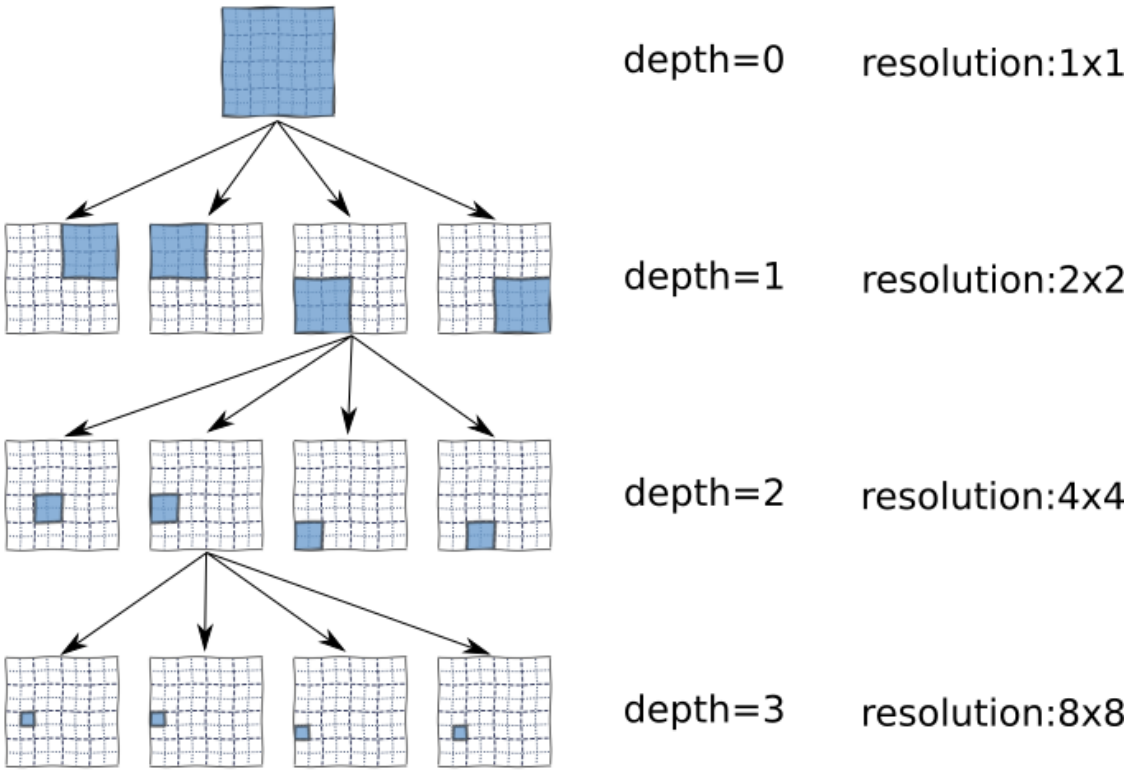


Figure 2.2: The box splitting of quadtree data structure. Each branch of the quadtree stores the location information within the corresponding boxes. The quadtree data structure resemble the box-counting method (which will be discussed in section2.3), that each depth represents a scale or resolution.

PR-Qtree uses a spatial decomposition strategy that split the area by inserting a dummy point (grey node) at the center of a cell, which splits the area into four equal size quadrants,

doi:10.6342/NTU201802404

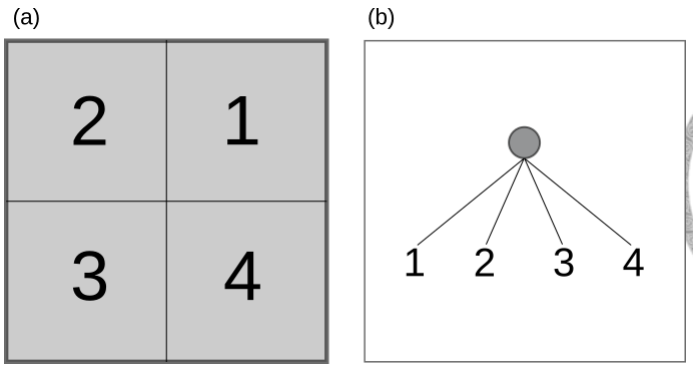


Figure 2.3: The branch-id assigning rule for indicating the (a) points locations and (b) branches positions.

namely box-splitting. The four equal size quadrants (child-nodes) were separated using the x- and y-coordinates (x_{grey} and y_{grey}) of the dummy point: any location in the upper-right cell (id-1 cell in Figure 2.3a) has a x-coordinate (x) larger than or equal to the x_{grey}, and a y-coordinate (y) larger than or equal to the y_{grey}; similarly, $x \leq x_{grey}$ and $y \geq y_{grey}$ in the upper-left (id-2), $x \leq x_{grey}$ and $y \leq y_{grey}$ in the lower-left (id-3), $x \geq x_{grey}$ and $y \leq y_{grey}$ in the lower-right (id-4). This box-splitting procedure is repeated at cells where there contains more than one point, until there are at most one point in each cell.

A demonstration of inserting three points into the PR-Qtree is shown in Figure 2.4. In the process of inserting the first point (Figure 2.4a), since there are no other points in the study area, no box-splitting is needed and one large cell is generated with the same area as the whole study area. The first point has then occupied the cell, i.e. the root node at the top level (depth-0, Figure 2.4e). Since the previous cell (root-node) already contains one point, the insertion of the second point forces the tree to split (Figure 2.4b). After the first box-splitting process, four boxes are generated, the two points are separated into different cells and occupied two different leaves at depth-1 in the quadtree (Figure 2.4f). To insert the third point, the second box-splitting process happens. After the second the box-splitting, the second and third points are still sharing a cell (Figure 2.4c) and sharing a leaf node (Figure 2.4g) due to the short distance between the two points. Therefore, the third box-splitting process occurs and which separates the second and third points into different cells (Figure 2.4d); the two points are now occupying different leaves at depth-3 (Figure 2.4h).

doi:10.6342/NTU201802404

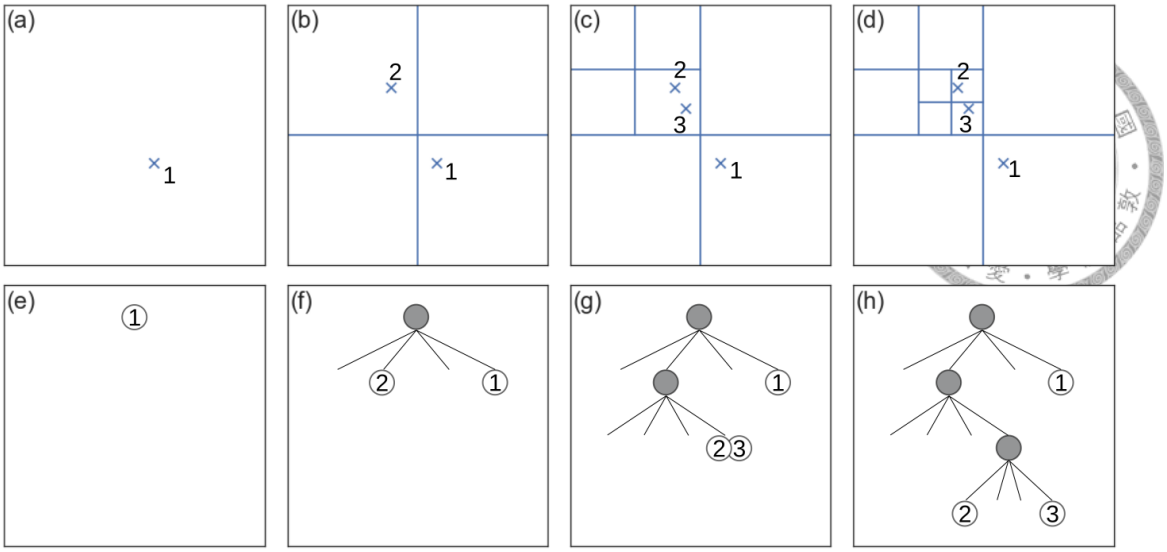
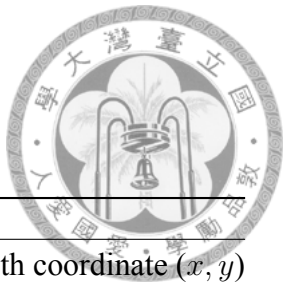


Figure 2.4: A demonstration of the generation of the quadtree. The subplots in the first row (a-d) show the insertion of three nodes and the box splitting process; the subplots in the second row (e-h) show the generation of the quadtree while the corresponding splitting process happened, respectively to the first row.

The idea and concept of splitting the study areas and generating the quadtree are shown and discussed in Figure 2.4. The corresponding pseudo-code for generating the quadtree is shown in Algorithm 1, which includes the point-insertion and box-splitting procedures.

In Figure 2.5, the John Snow cholera point events (Arribas-Bel et al., 2017) were used to demonstrate the PR-Qtree generation and the resulting quadrant cells. Each of the events was separated into a cell. Some of the events were located in smaller cells, while some others occupied larger cells. This was because the points distribution showed a clustering pattern, that some points were near to each other, forcing the PR-Qtree to split more times to separate the points into distinct cells.

doi:10.6342/NTU201802404



Algorithm 1 Point Region Quadtree

procedure Point-Insertion($P(x, y)$) ▷ a point with coordinate (x, y)
  if $root$ is None **then**
    $root \leftarrow P(x, y)$ ▷ the point inserted at root position
  else
    $this\_node \leftarrow root$
    while typeof($this\_node$) $\neq leaf$ **do** ▷ until get the leaf node
      compare P(x,y) with $this\_node$
      if P(x,y) in NE quadrant **then**
        $this\_node \leftarrow this\_node.NE$
      else if P(x,y) in NW quadrant **then**
        $this\_node \leftarrow this\_node.NW$
      else if P(x,y) in SW quadrant **then**
        $this\_node \leftarrow this\_node.SW$
      else
        $this\_node \leftarrow this\_node.SE$
      end if
    end while ▷ $this\_node$ is now a leaf node
    if $this\_node.point \neq None$ **then** ▷ $this\_node$ is occupied
      $Q(x, y) \leftarrow this\_node.point$
      Box-Splitting($this\_node$)
      Point-Insertion($Q(x, y)$)
      Point-Insertion($P(x, y)$)
    else
      $this\_node.point \leftarrow P(x, y)$ ▷ the point is inserted here
    end if
  end if ▷ done insert point
end procedure
procedure Box-Splitting($this\_node$)
  $x_0, y_0, x_1, y_1 \leftarrow this\_node.box$
  $x_c \leftarrow (x_1 - x_0)/2. + x_0$ ▷ split the box at the center
  $y_c \leftarrow (y_1 - y_0)/2. + y_0$
  $this\_node \leftarrow C(x_c, y_c)$
  typeof($this\_node.childs$) $\leftarrow leaf$ ▷ childs include NE, NW, SW, SE
  typeof($this\_node$) $\leftarrow grey$
end procedure ▷ done splitting box

doi:10.6342/NTU201802404

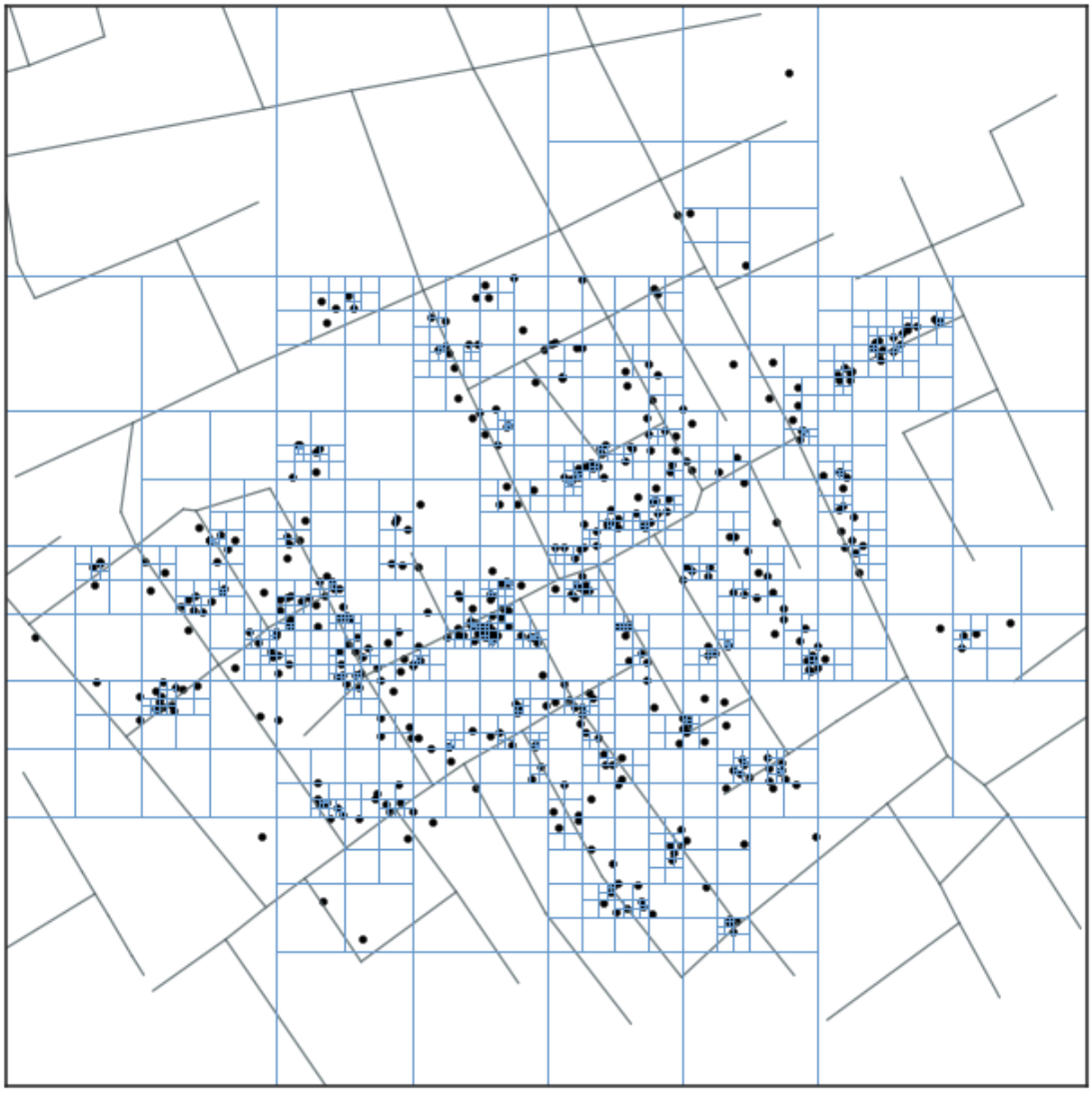
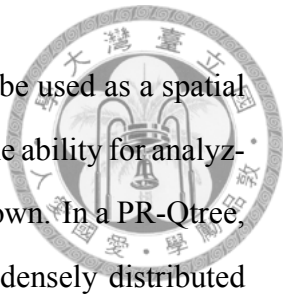


Figure 2.5: The point distribution and generated PR-Qtree based on the John Snow's cholera point data (Arribas-Bel et al., 2017).

doi:10.6342/NTU201802404

2.2 Leveling-Down



The PR-Qtree was designed not to be used as an analysis tool, but to be used as a spatial indexing data structure. This study modified the PR-Qtree to include the ability for analyzing the scaling properties, by adding a procedure named as leveling-down. In a PR-Qtree, some data points reach to a deeper level of the tree because of the densely distributed points (e.g. the second and third points in the previous demonstration), while some points stop at a shallower level of the tree because of the sparsely distributed points (e.g. the first point). Therefore, **the points that stop at a shallower level would be ignored while analyzing the deeper levels**, i.e. the larger boxes will not be counted at a deeper scale.

For the previous example in Figure 2.4, in Figure 2.6a, we are counting the number of boxes that is occupied by at least one point (occupied box or occupied branch). All points are considered in the same box at depth-0, therefore occupied box equals one. At depth-2, two points are located in the north-west quadrant (second branch) and one in the south-east quadrant (fourth branch). In these two scales, all three points are considered and contributed to the calculation of occupied box and ratio. One point stops at depth-1, leaving only two points are continuing the development of the tree, i.e. at depth-3 and depth-4. Hence, the level-based measurements, that is used to compare the properties of each level, could be problematic because the proportion of points reached different levels do not equal. In other words, to archive the goal to compare the distribution characteristics in different levels (which represent different scales), the tree should be converted so that the proportion of the covered points in each level is same and equal to all of the points.

In simple words, the leveling down process drags the points that stop before reaching the maximum depth of the tree down to the maximum depth. Take the three points in Figure 2.4 for example, the point-1 stops at depth-1 while the maximum depth of the tree is 3. The point-1 is then dragged down to the depth-3 by splitting the cell two more times, which process leads to the generation of two grey nodes at depth-1 and depth-2 and put in the corresponding position in depth-3. The resulting boxes and tree structure are shown in Figure 2.7.

doi:10.6342/NTU201802404

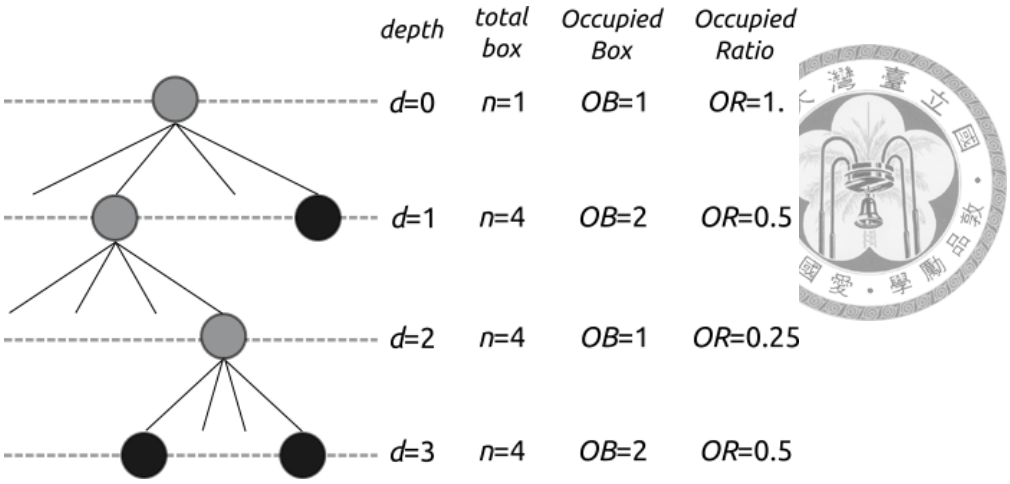


Figure 2.6: The box counts in each depth (scale) for the example in Figure 2.4. Due to the situation that points stopped at different depth, the basis to compare different depths is using different number of points.

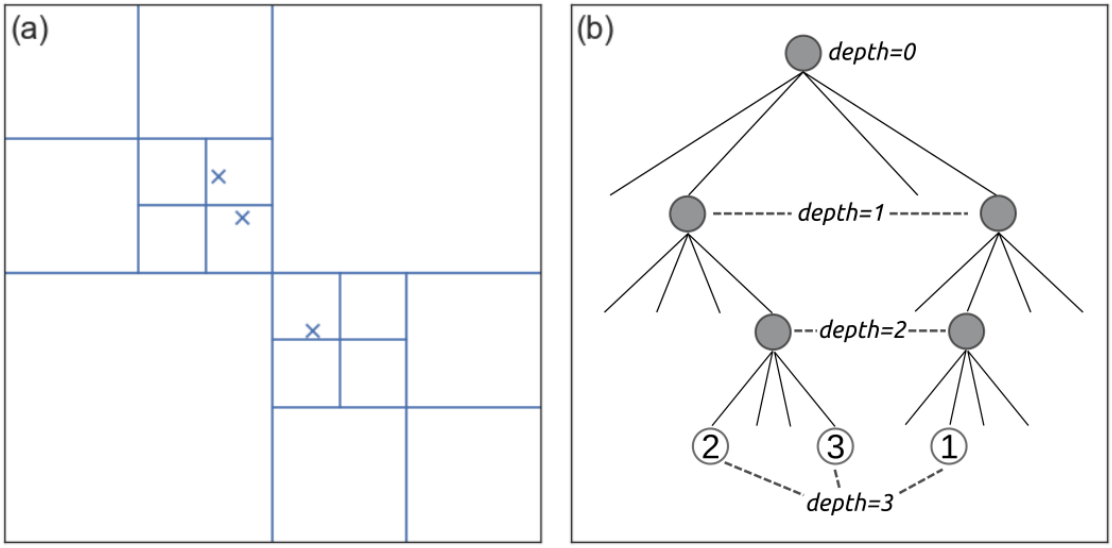


Figure 2.7: The leveling-down version of boxes (a) and quadtree (b) structures using the previous example. The points located at the box with the same size in (a). The branches are developed until the deepest levels so that the points are at the leaf of the last depth, and the development process of the points left a trail of grey nodes in the levels before the last depth (b).

To demonstrate the changes in the tree structure, the John Snow cholera data(Arribas-Bel et al., 2017) is used again in the following discussion. In Figure 2.8a, which shows the result before leveling down procedure, the boxes that are occupied by at least one point, namely occupied box, is counted for each depth (i.e. the number of gray nodes in each depth). The counted boxes of the depth-7 are lower than depth-6. This is because most points stopped at depth-6, therefore fewer points showed after depth-6, and the counting in

doi:10.6342/NTU201802404

depth-7 do not include the distribution of points that stopped before depth-7. Figure 2.8b, the occupied box number shows an increasing trend, this is because the points that previously stopped at shallower depth are now contributing one occupied box at the levels after where it previously stopped (same as the point-1 in Figure 2.7 at depth-1 and depth-2). In other words, the leveling down procedure makes sure all points appear in the calculation of all depths; or in other words, all points are considered in the calculation of all scales.

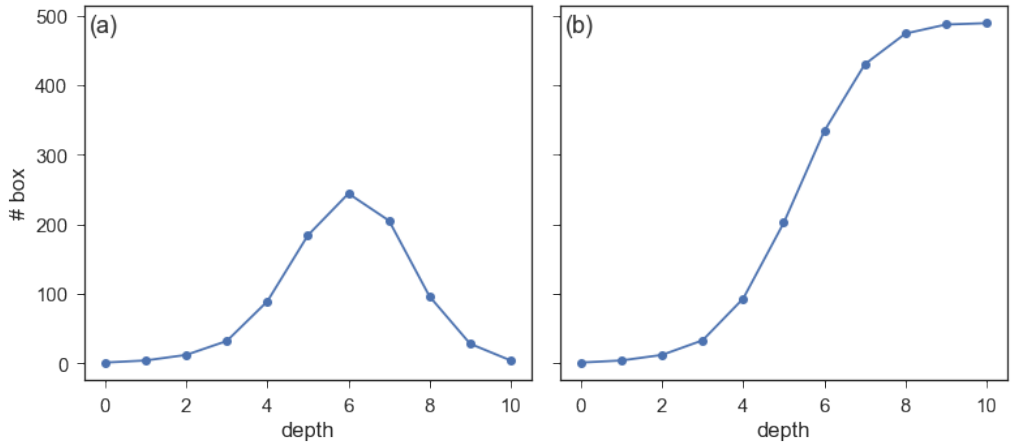


Figure 2.8: The occupied box by depth (scale) of (a) before leveling down procedure and (b) after leveling down procedure, for the John Snow cholera data (Arribas-Bel et al., 2017).

Figure 2.9 shows the changes in boxes structure before and after the leveling down procedure. The non-leveling down version is shown on the left, in which some of the points occupied a larger cell, and some of them are within a smaller cell. On the right, in the leveling down version, each of the points occupied the smallest size box, even when its parent box (cell at one upper level) do not have other child boxes (sibling cell).

The algorithm of leveling-down is shown in Algorithm 2. The number of running the dragging process depends on where each point stopped, and the number of levels left to reach the deepest level. The dragging process used two previously defined procedures (Box-Splitting and Point-Insertion), i.e. forcing the node to split and replace the point with a newly generated grey node; then reinsert the point to the tree.

After the construction of PR-Qtree and the leveling-down procedures, the points distribution is converted into the tree structure that contains the information of the points distribution in different scales (depths). Theoretically speaking, **a PR-Qtree represents**

doi:10.6342/NTU201802404

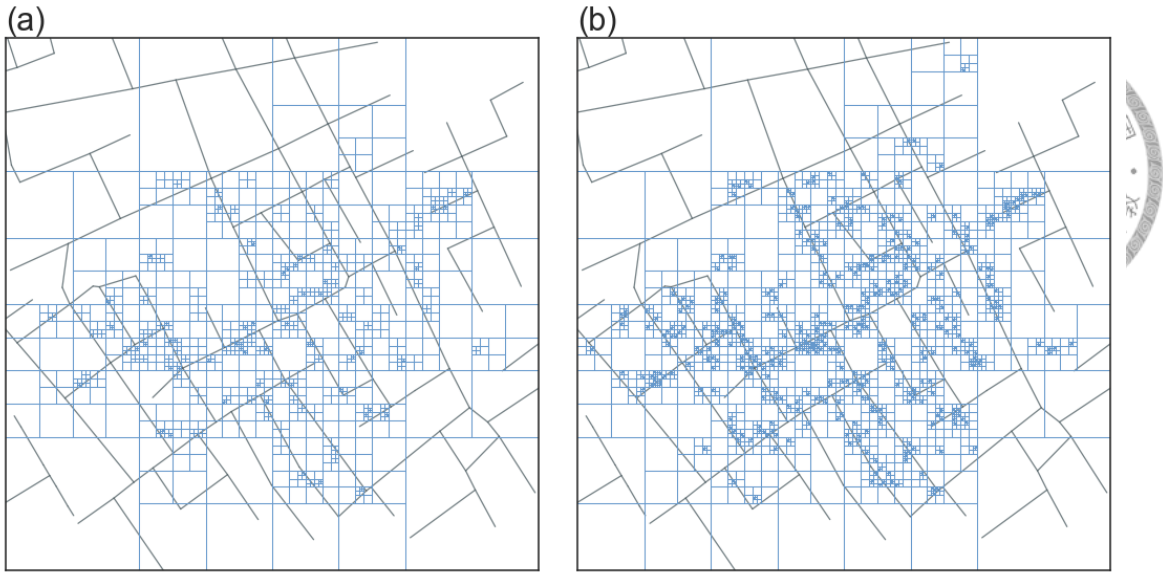


Figure 2.9: An demonstration of the PR-Qtree based on the John Snow's cholera point data (Arribas-Bel et al., 2017): (a) the boxes before leveling down, and (b) after leveling down.

Algorithm 2 Leveling-down

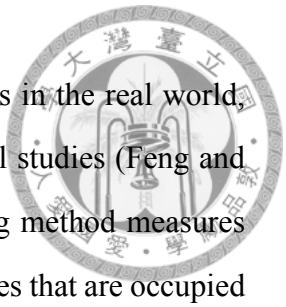
```
procedure Leveling-down(Tree)
    d_max ← max([p.depth for p in Tree.points])
    for p in Tree.points do
        if p.depth ≤ (d_max − 1) then
            for (i=0; i < (d_max − p.depth); i+=1) do
                Box-Splitting(p.node)          ▷ in algorithm 1
                Point-Insertion(p)             ▷ in algorithm 1
            end for
        end if
    end for                                    ▷ done leveling down all points in Tree
end procedure
```

the corresponding point distribution. The existence of a grey node on a position somewhere within the tree indicated that the area of the box contains at least a point. The grey node at a higher level of the tree (i.e. lower depth number) indicated the box area of the grey node is larger, e.g. the box size of a grey node at depth 1 is one quarter of the whole study area, and the box size of a grey node at depth 2 is one eighth of the study area. In other words, **the depth of the PR-Qtree can also be thought as the scales of different size of resolution**.

The third and fourth procedures (the following two subsections) were designed to analyze the PR-Qtree structure for extracting the scaling properties of the points distribution.

doi:10.6342/NTU201802404

2.3 Box-counting method



To empirically measure the fractal geometry of geographical features in the real world, the box-counting method was introduced and applied to geographical studies (Feng and Chen, 2010; Jiang and Liu, 2012; Frankhauser, 2015). Box-counting method measures the fractal geometry in terms of quantifying the boxes in different scales that are occupied by the geographical features. The basic box-counting method and calculation concept are discussed in this section.

The concept of box-counting method can be captured by the procedures of PR-Qtree (Figure 2.10). Given a square-shape study area, the box-counting method starts by splitting the study area using several large and equal size box. The John Snow data is used as an example in Figure 2.11 for demonstration. While using the largest size boxes as in (a), all four boxes are occupied; a second largest boxes in (b), a total of 16 boxes are generated, only 12 of them are occupied; using third largest size boxes to split the study area (c), 33 out of 64 boxes are occupied. This box-splitting and counting occupied procedures keep going until the number of occupied boxes reaches the maximum value (i.e. the total number of points).

According to the box-counting method (Frankhauser, 2004, 2015), the fractal dimension of the distribution of a geographical features can be calculated through fitting the slope of the linear curve in a $logN$ - $logr$ plot (Figure 2.12a), where N and r are the number of occupied boxes and the ratio of side length to the largest side length of the box in each step. In other words, as the size of boxes reduced (the resolution increased, and the scale becomes higher and finer), there should be more boxes overlap with the geographical features (e.g. built environment or the point locations of events), hence the higher number of occupied boxes. If the distribution of the geographical features meets the fractal properties, the correlation between the number of occupied boxes and the side length would be strong, and an obvious and straight fitting line would exist (the part where the side length ratio is relatively larger in Figure 2.12a). Based on the calculation using the side length, i.e. the way of capturing the scaling properties, the relationship is expected to be negatively correlated, so the calculation of the fractal dimension has then multiplied the

doi:10.6342/NTU201802404

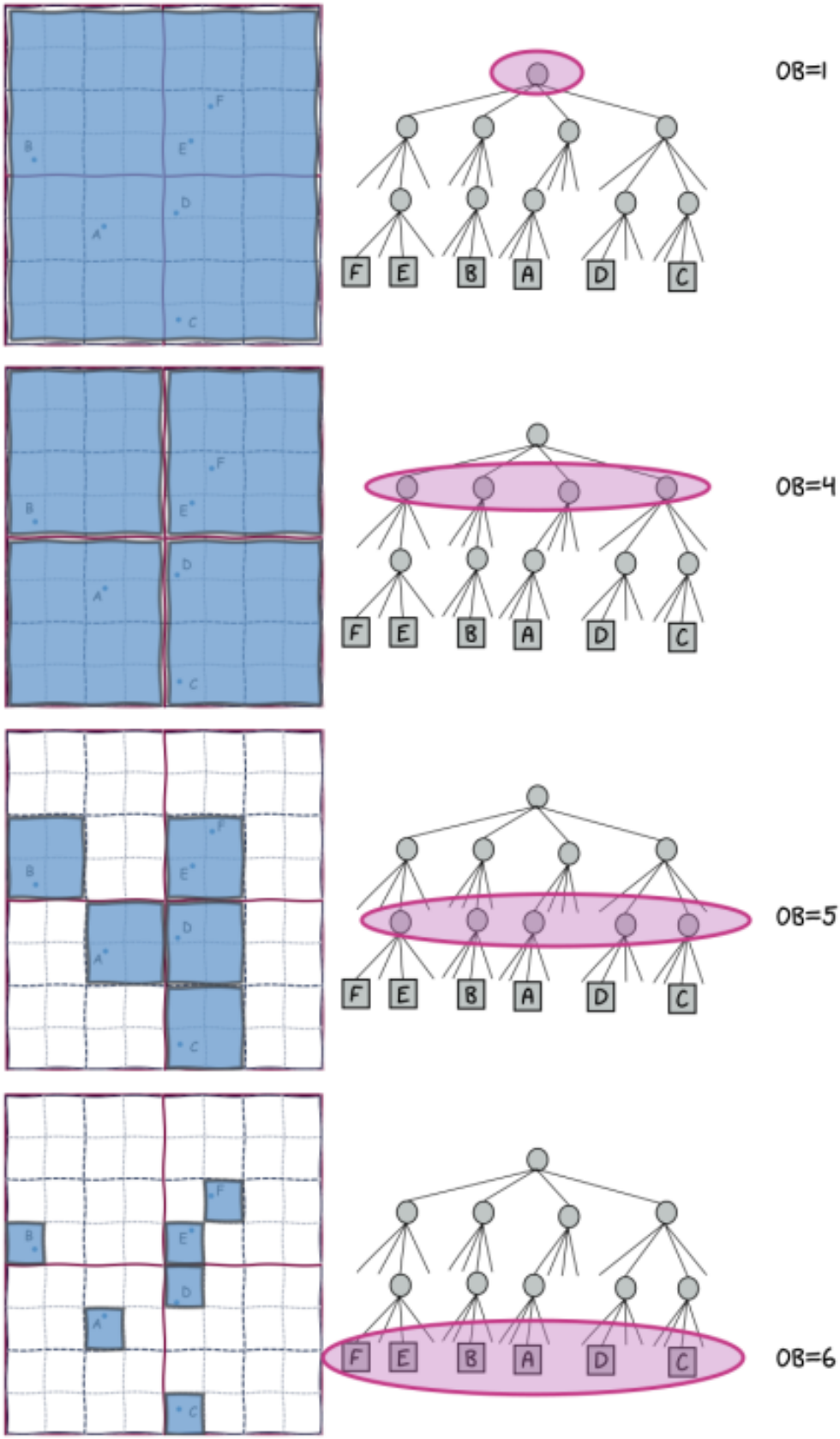


Figure 2.10: The box counting procedure using quadtree environment.

doi:10.6342/NTU201802404

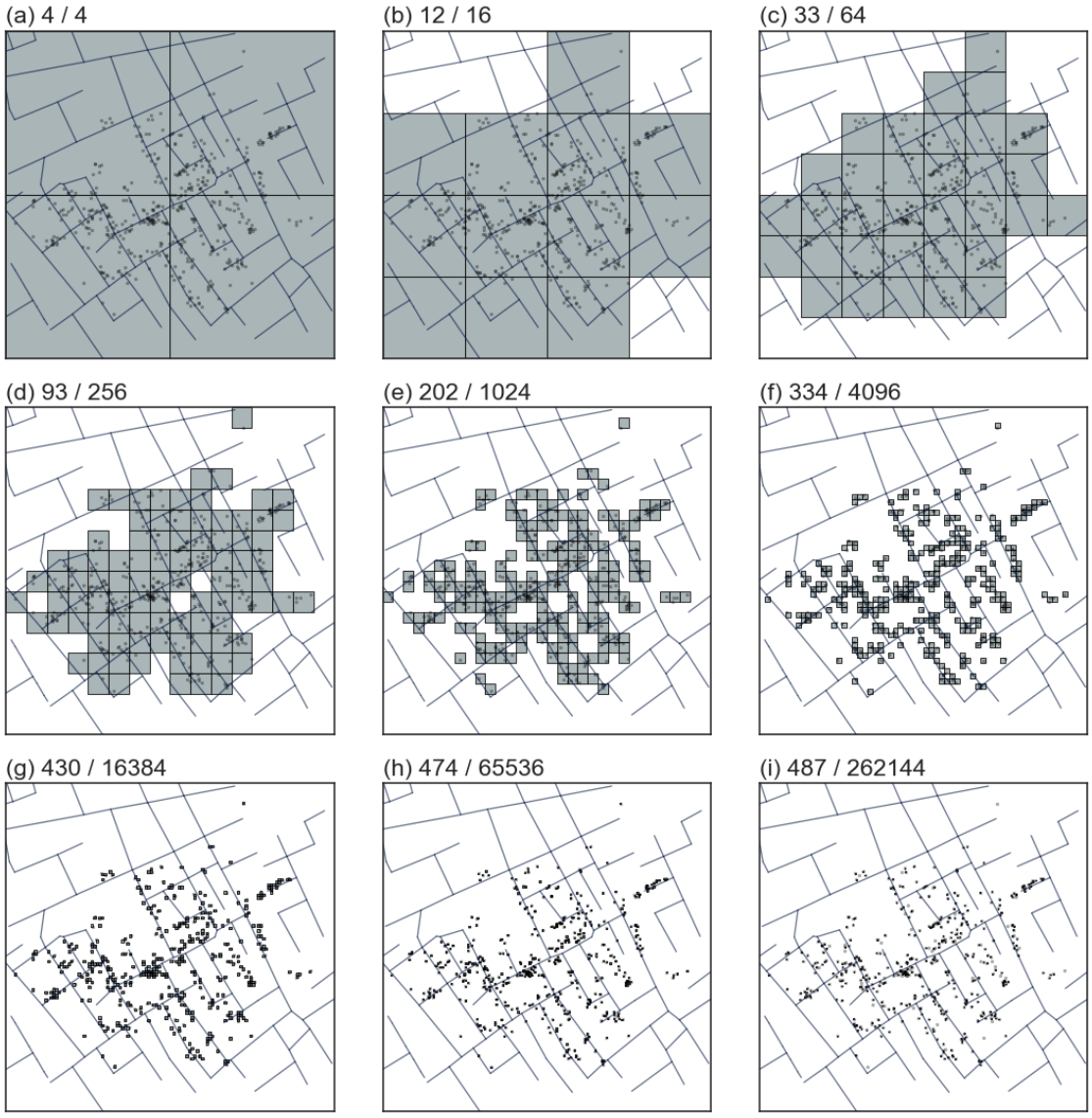


Figure 2.11: An demonstration of the box-counting method based on the John Snow's cholera point data (Arribas-Bel et al., 2017). The ratio above each sub-figure indicates the number of occupied box divided by the total number of box in each scale.

slope of the fitting line with -1 to calculate the fractal dimension of the distribution (Equation 2.1). In a two dimensions space, the maximum fractal dimension is 2, which is the same dimension of a rectangle shape. Therefore, if the fractal dimension of a distribution is approximately 2, this indicated that the distribution of the geographical features fill up almost all of the study area; every time a box is divided into four sub-boxes ($l_{i+1} = l_i \times 2^{-1}$, hence, $r_{i+1} = r_i \times 2^{-1}$), the four sub-boxes are all occupied ($N_{i+1} = N_i \times 2^2$), and this situation continues in all steps (scales).

doi:10.6342/NTU201802404

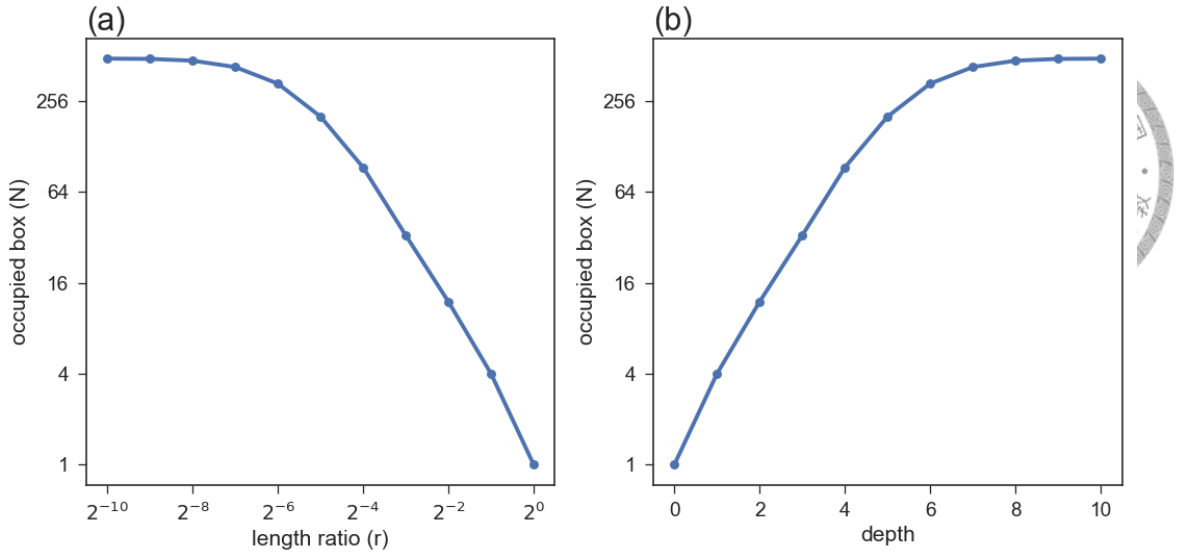


Figure 2.12: The occupied box count (a) by the ratio of side length, and (b) by the depth (scale). In (a) the larger the ratio of side length, the larger the box, thus the number of the occupied box is lower, starting from one. The fractal dimension occurs on the right part where the r is larger and forms a fitted line with a negative slope. In (b) the shallower depth (smaller depth value) indicated larger box, therefore the shape of the line (occupied box) shows opposite pattern in the two sub-figures, which represents the same pattern. Thus, the fractal dimension can be calculated using the depth value and the log number of the occupied box as a positive value.

$$r = \frac{l}{L} \tag{2.1.1}$$

$$N = r^{-FD} \tag{2.1.2}$$

$$log_2 N = -FD \times log_2 r \tag{2.1.3}$$

$$FD = -\frac{log_2 N}{log_2 r} \tag{2.1.4}$$

where, FD is the fractal dimension, N is the number of occupied boxes in each step, l is the side length of the box in each step, L is the side length of the largest square area, and R is the ratio of how long the side length of a step is compared to the whole side.

doi:10.6342/NTU201802404

The calculation of the fractal dimension can be simplified in the PR-Qtree structure. The side length of each cell in a level of depth can be calculated by dividing the whole side length by 2 with an exponent of the depth level (Equation 2.2). Thus, the negative logarithm of the ratio of the side length to the total side length is equal to the depth value (Figure 2.12b), hence can be replaced by the negative depth value; and the equation of fractal dimension (Equation 2.1) can be simplified to the Equation 2.3. There is no negative sign in this equation, indicating that the slope of the two variables should be positive, i.e. the higher the depth level (the deeper the tree), the higher the log number of occupied boxes. Since this fractal dimension is calculated for the scaling range starting from the largest box size, the fractal dimension indicated the fractal properties of the point pattern as a whole, i.e. the global fractal properties. Therefore, in this study this fractal dimension is named global fractal (Equation 2.3).

$$l_{depth} = \frac{L}{2^{depth}} \tag{2.2.1}$$

$$2^{-depth} = \frac{l_{depth}}{L} \tag{2.2.2}$$

$$log_2 2^{-depth} = log_2 \frac{l_{depth}}{L} \tag{2.2.3}$$

$$-depth = log_2 \frac{l_{depth}}{L} \tag{2.2.4}$$

$$-depth = log_2 r \tag{2.2.5}$$

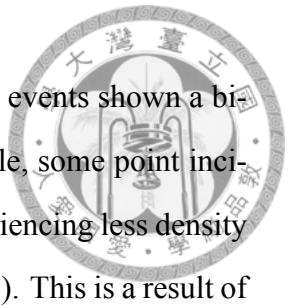
where, $depth$ is the depth level of the step, which is started from 0 at the root level.

$$Global\_fractal = \frac{log_2 OB}{depth} \tag{2.3}$$

where OB represents the number of occupied boxes.

doi:10.6342/NTU201802404

2.4 Searching for critical scale



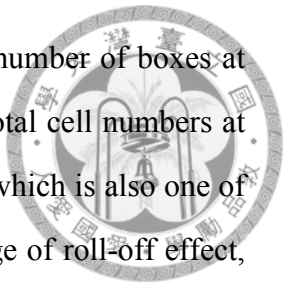
Previous studies indicated that the distributions of geographical point events shown a bi-fractal pattern (Agterberg, 2013; Chen and Wang, 2013). For example, some point incidents may happen more frequently in the core area of a city, and experiencing less density at the periphery areas (Tannier and Pumain, 2005; Thomas et al., 2008). This is a result of the heterogeneity nature of the distribution of geographical features (Goodchild and Mark, 1987; Batty, 2008). **Therefore, a turning point of scale (assuming bi-fractal) must exist in between the scale range.** For example, Agterberg (2013) used the first eight points to calculate the fractal dimension; the scales (with smaller boxes) were experiencing the so-called roll-off effect (Pickering et al., 1995; Walsh et al., 1991).

Conceptually speaking, the **turning point of scale** indicates that the range of scales before the turning point experience the global fractal properties, that is the distribution can be explained by a fractal pattern as discussed in the previous sub-section (if the fractal properties exist). The range of scales after the turning point experience the roll-off effect, that the distribution is beyond the explanation of the primary fractal pattern (the overall fractal properties, treating the distribution as a whole); they could be a local fractal pattern that happens only at some parts of the study area. In this study, this turning point of scale is called the critical scale, because it represents the optimized and finest scales that can be explained by the global fractal pattern, and starting to turn into the local fractal pattern after this turning point.

In order to calculate the local fractal dimension, this study introduces an index called the occupied ratio (Equation 2.4). The occupied ratio converts the occupied box into a proportion by dividing the count of boxes by the total number of cells in each level. After a series of box-splitting processes, the number of all boxes (occupied and non-occupied) should equal to four with an exponent of the depth level. By dividing the count of occupied boxes by the total number of cells in each depth will convert the value into a 0.0 to 1.0 floating number that represents the coverage ratio (occupied ratio).

$$OR_i = \frac{OB_i}{4^{depth}} \tag{2.4}$$

doi:10.6342/NTU201802404



While the occupied box count emphasizes the increment of the number of boxes at the earlier stage, the occupied ratio emphasizes the changes of the total cell numbers at the later stage, where the increment of occupied boxes is very slow, which is also one of the reasons the scales are experiencing the roll-off effect. In the stage of roll-off effect, increasing a level will create more cells, but the increment of the occupied box count is relatively small, which leads to the decreasing trend for the occupied ratio. Therefore, the log occupied ratio should be negatively correlated to the depth (or negative log ratio of side length). Thus, the $Local\_fractal$ can be calculated using Equation 2.5, by fitting a line to the relationships between depth and occupied ratio. The negative of the slope value is the $Local_fractal$. The number of total cell number increased by 4 times larger than the previous total cell number. Therefore, conceptually speaking, the largest possible value of the second dimension is less than 2, where the numerator (number of the occupied boxes) is very small and increases very slowly in comparison to the denominator (total number of cells).

$$local\_fractal = -\frac{log_2 OR}{depth} \tag{2.5}$$

The calculation of the global factor should consider only the scales before the turning point of scale, whereas the calculation of the local factor should consider only the scales after the turning point. The **objective of the optimization model is to search for an optimal resolution (scale)**, i.e. the turning point of scale, so that the global (primary) and local (secondary) fractal dimensions can be computed using the optimized turning point as a cutoff.

Figure 2.13 shows a demonstration of the optimization model for searching the turning point, i.e. the critical scale. The model will try to fit both the global (Figure 2.13a) and local (Figure 2.13b) scaling lines (the green line in (a) that shows the occupied box, and red line in (b) that shows the occupied ratio) using the same cutoff, which is all of the possible depth value. The dashed blue lines in the two sub-figures (a and b) show the

doi:10.6342/NTU201802404

fitted line. The possible depth values include the depth starting from 1 to the depth level before the max-depth. The Pearson correlation coefficients between the fitted lines and data for both global and local fractal dimensions (converted to R^2 as the green and red lines in Figure 2.13c) are multiplied to compute the fitness (objective value, blue line in Figure 2.13c), which is then used for maximization. In other words, the model tries to slide the cutoff value (turning point of scales) concurrently in both of the figures a and b, and calculate the fitting lines, then maximizing the two correlation results concurrently. The optimization model is described in Algorithm 3.

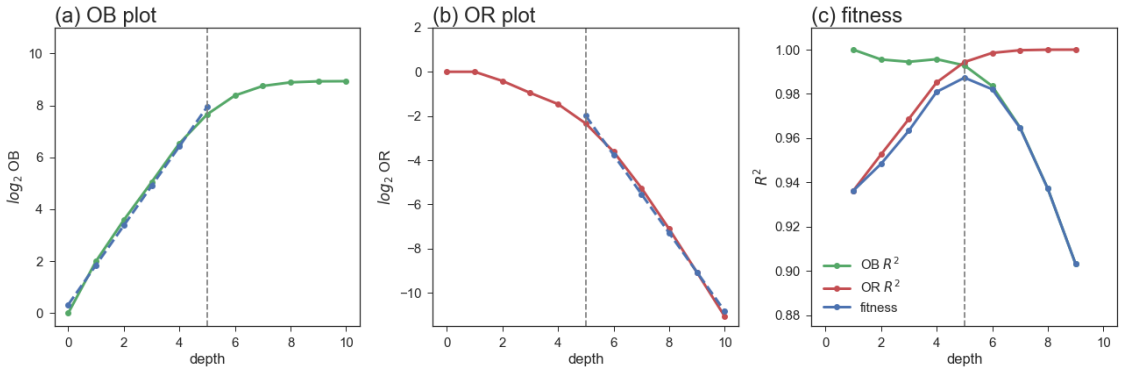


Figure 2.13: The optimization procedure tested each depth (scale) of the PR-Qtree as a cutoff point, and trying to fit a line before the cutoff point using the OB-plot (a), and fit a line after the cutoff point using the OR-plot (b), the coefficient of determination (R^2) of the two lines to the observed data (occupied box and occupied ratio, respectively) were shown in (c), together with the fitness. The result from this demonstration data (John Snow cholera data) shows that the critical scale is 5

Given a set of points, starting from the construction of PR-Qtree and leveling down procedure, to the calculation of global and local fractal dimension, and finally, the optimization model, **the critical scale of the given point distribution will be identified**. For the John Snow cholera data as shown in previous sections, the critical scale is five, which is shown in Figure 2.11e.

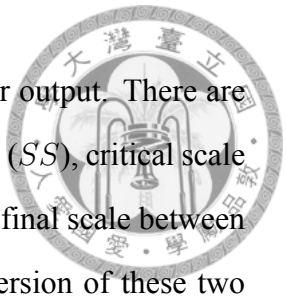
doi:10.6342/NTU201802404

Algorithm 3 Optimize-Critical Scale model

procedure Search-Critical-Scale($depth, OB, OR$) ▷ 3 vectors
  $FS \leftarrow max(depth)$
  SSgetMin(OB)
  $max\_fitness \leftarrow -9$ ▷ so it will be replaced immediately
  $CS \leftarrow None$
  $FD'_1, FD'_2 \leftarrow None, None$
  for (i=$SS + 1$; $i \leq (FS - 1)$; i+=1) **do** ▷ try each depth
    $FD_1, r_1^2 \leftarrow$ Fitting($depth[SS : i], log_2(OB)[SS : i]$)
    $-FD_2, r_2^2 \leftarrow$ Fitting($depth[i : FS], log_2(OR)[i : FS]$)
    $fitness \leftarrow r_1^2 \times r_2^2$
    if $fitness > max\_fitness$ **then**
      $SS \leftarrow i$
      $max\_fitness \leftarrow fitness$
      $FD'_1 \leftarrow FD_1$
      $FD'_2 \leftarrow FD_2$
    end if
  end for ▷ done searching
  return SS, CS, FS
end procedure
procedure Fitting(X, Y)
  $fitted\_curve(X, Y) \leftarrow$ with first degree polynomial equation
  get $slope$ from $fitted\_curve$
  compute $E(Y)$ vector from $fitted\_curve$
  $r \leftarrow Pearson\_correlation(Y, E(Y))$
  return $slope, r^2$
end procedure
procedure getMin(OB)
  $d_{min} = 0$
  for (i=0; $i \leq (d_{max} - 1)$; i+=1) **do**
    $y_1 = OB[i]$
    $y_2 = OB[i + 1]$
    if $y_1 == y_2$ **then**
      $d_{min} \leftarrow i + 1$ ▷ replace SS
    end if
    if $y_2 > 2$ **then**
      break ▷ break the loop
    end if
  end for
  return d_{min}
end procedure

doi:10.6342/NTU201802404

2.5 Preparation of the results



The final step of the analyzing framework is to prepare the results for output. There are three key results based on the Algorithm 3, including the starting scale (SS), critical scale (CS), and final scale (FS). In order to compare the critical scale and final scale between different datasets with a different number of points, a normalized version of these two scales were introduced and discussed. On the other hand, the relative critical scale is also introduced to present the proportion of macro view and micro view of distribution. Therefore, a total of six indexes will be calculated as the output of the framework, namely the scaling properties.

In addition, based on the PR-Qtree, this study also proposes a model to aggregate points into different scales (depth in the PR-Qtree). The point aggregation model is described following the six indexes.

2.5.1 The key scales of distribution

Starting scale

While the point distribution is concentrated only at a small part of the study area, the occupied box plot will stop at the beginning (shallower depth), and rise again after the beginning range. Therefore, this study defined the scale when the scaling started as the starting scale (SS). The detection of starting scale is described as a function ($getMin$) in Algorithm 3.

For example in Figure 2.14, the point distribution in Figure 2.14a shows a cluster locates at the left corner, thus the number of the occupied box in depth-1 (dashed lines) is equal to depth-0, which is one box. As shown in Figure 2.14b, this situation leads to the observed line in OB-plot (blue line) becomes horizontal at the beginning (because the log OB is both $log_2 1 = 0$). There is no fractal pattern from the whole study area because the other three boxes (top left, top right, and bottom right) are all empty, and which is different from the spatial distribution in the bottom left box, i.e. no self-similarity condition happens from the whole setup. The fractal pattern (self-similarity) exists while

doi:10.6342/NTU201802404

the study area is concentrated into the bottom left box, which self-similarity condition is shown in Figure 2.14b after depth-1. Therefore, we define the starting scale as which scale the fractal patterns starting to exist, i.e. depth-1 in this example.

The similar situation happens for the distribution in Figure 2.14c, that the cluster concentrates at the center of the study area. This leads to the situation that the occupied box is both four for the depth-1 and depth-2, and forms a stopping step before depth-2 (Figure 2.14d). As shown in Figure 2.14c, the outer ring (which were split in depth-2, dotted lines) are empty, and the four boxes at the center area are occupied. This pattern is not the fractal pattern of the point distribution, which only occurs at the first scale; the fractal pattern exists after we zoom into the four occupied boxes at the center. Therefore, the starting scale of this example is equal to depth-2 (Figure 2.14d).

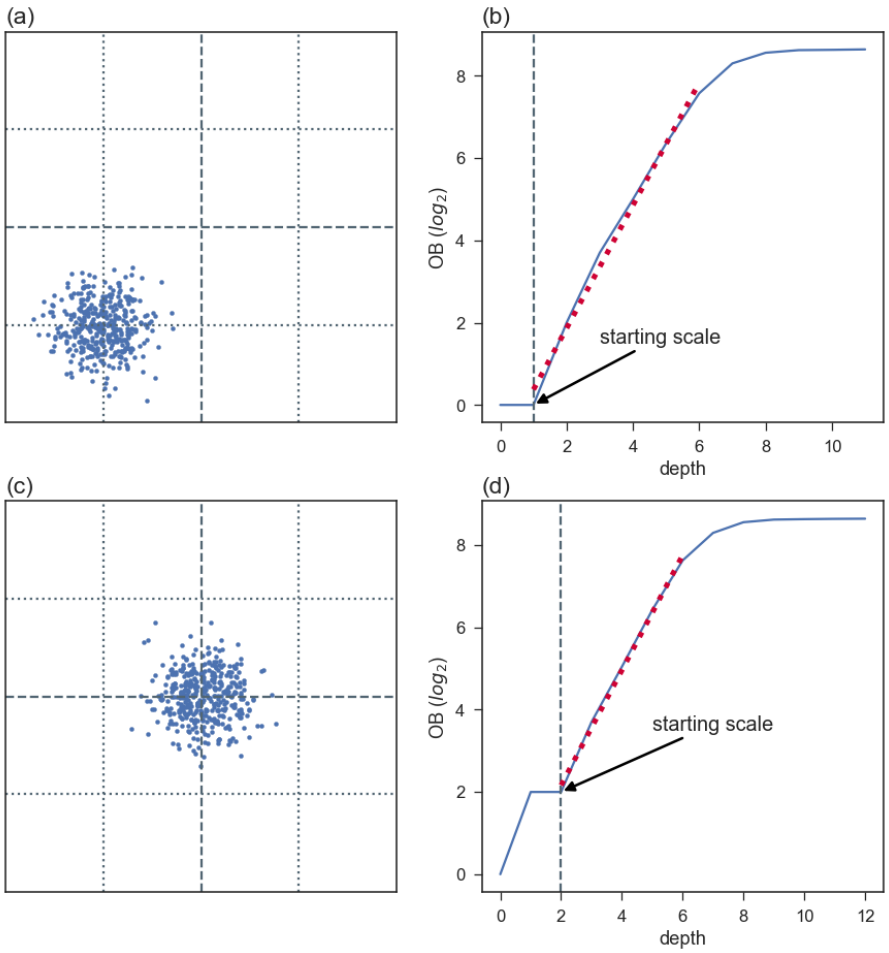


Figure 2.14: Two example of distribution with starting scales>0 (a) and (c), and their corresponding OB-plot (b and d).

doi:10.6342/NTU201802404

Critical scale

Critical scale (CS) is defined and discussed in Section 2.4. In summary, critical scale represents the boundary which cut off the scales that shows a fractal property, i.e. distribution that is self-similar in all scales within the range, from the scales that is under the roll-off effect. Figure 2.11 shows the distribution of occupied boxes in different scales (depth). Note that the patterns show some similar properties between depth-2 to depth-5 while zooming in at the higher resolution, indicating the self-similarity properties exists and which is scale-invariant. While zooming into the map of depth-6 and depth-7, there are some places shows a similar clustered (as above), but there are also some places with no occupied box within the zooming window. In other words, the critical scale is the finest resolution that can be explained by the scaling properties, with the corresponding global scaling pattern ($Global\_fractal$). On the other side of the critical scale, there are the scales that some micro and small clusters existed, and which scale-range forms the local scaling pattern ($Local\_fractal$).

Final scale

The Final scale (FS), i.e. the maximum depth of the PR-Qtree, represents the finest scale that is needed to separate each point into a single cell. If the points are regularly spaced and placed (Figure 2.15a), the constructed PR-Qtree will be very balanced, that all of the leaves located in the lowest or the second lowest nodes before the leveling-down process (Figure 2.15b). On the other hand, the number of empty leaves is also low due to the balance tree properties. In this condition, the final scale of the PR-Qtree is therefore shallower. In contrast, if the same number of points are clustered in a small area (Figure 2.15c), that some of the points are very close to each other, the PR-Qtree will be unbalanced, that some branches are stopped at a shallower depth, and some branches are heavily grown (Figure 2.15d). Note that there are many empty leaves at different scales since the corresponding cells are empty. This condition will lead to a deeper final scale with the same number of points, i.e. the finest cell is smaller in comparison to the regular case.

doi:10.6342/NTU201802404

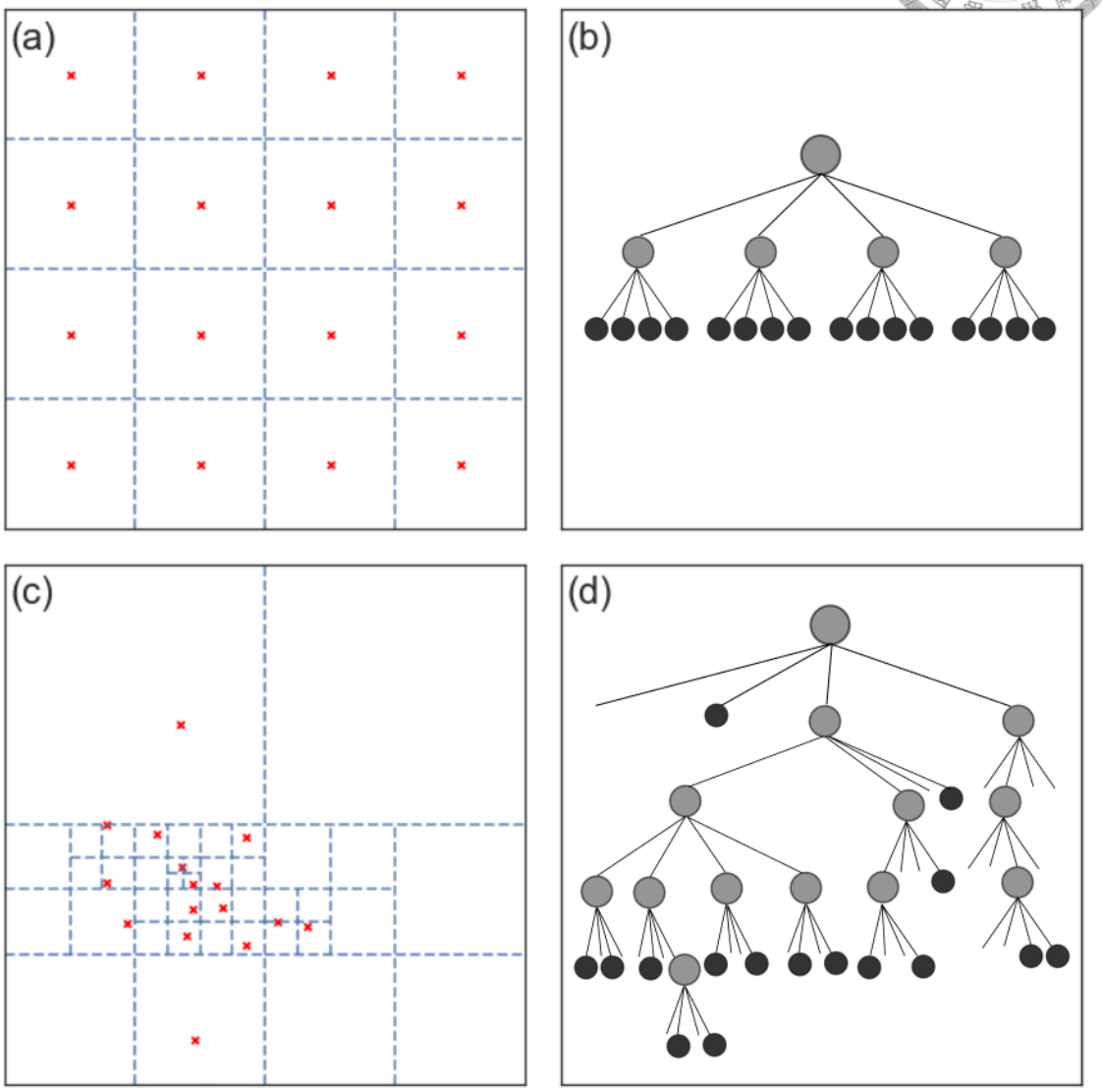


Figure 2.15: The maximum depth of two types of distribution with the same number of points ($n = 16$): (a) completely regular points distribution, and (b) the corresponding PR-Qtree for (a) before the leveling-down process; (c) clustered points distribution, and (d) the corresponding PR-Qtree for (c) before the leveling-down process. The maximum depth (i.e. FS) of the tree in (b) is 2, whereas the maximum depth of the tree in (d) is 5.

doi:10.6342/NTU201802404

2.5.2 Additional indexes of scales

Final scaling magnitude

One of the factors that will influence the final scale of the distribution is the number of points. For example, Figure 2.16 shows the influence of the number of nodes to the FS (and CS, which will be discussed in the next part). The distribution with more number of nodes will inevitably result in deeper FS.

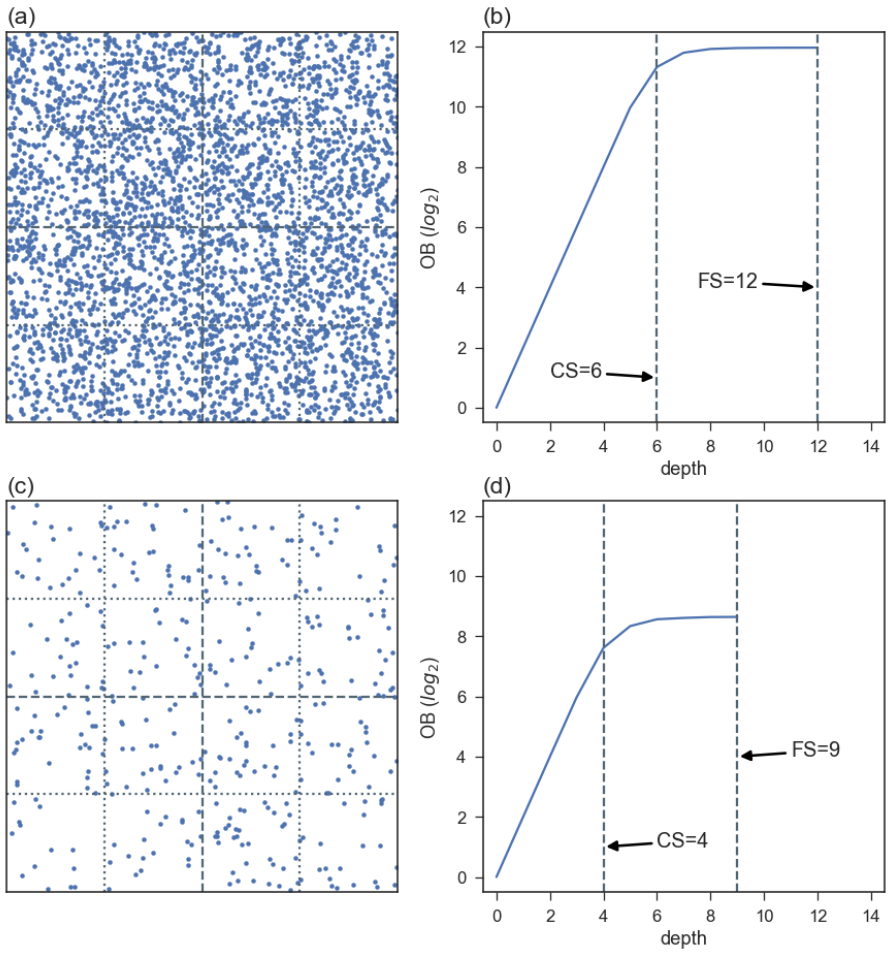


Figure 2.16: The critical and final scales with the varying number of points. The example in the top row shows: (a) the random distribution with 4000 points, and (b) the OB-plot with the annotation indicating CS and FS. The example in the bottom row shows: (c) the random distribution with 400 points, and (d) the corresponding OB-plot that indicates the CS and FS.

doi:10.6342/NTU201802404

Therefore, if two sets of point distribution with a different number of points (varying data size) are needed to be compared, a normalization version of FS is needed. This study proposes a normalization measurement for FS, namely final scaling magnitude (FS'). The concept of final scaling magnitude is to compare the FS of observed data with the FS of a completely regular point distribution with the same data size. As discussed in Figure 2.15, given a fix data size, a regularly distributed points will have the theoretically lowest FS. Thus, this study defines the final scaling magnitude as of how large is the observed FS in comparison to the regular distributed FS. While the $FS_{regular}$ is the smallest possible value of FS for a given number of points, FS' is a floating value greater than or equal to one. The calculation of FS' is shown in Equation 2.6.

$$FS' = \frac{FS_{observed}}{FS_{regular}} \tag{2.6.1}$$

$$FS_{regular} = ceil(\frac{log_2 N}{2}) \tag{2.6.2}$$

The regularly distributed $FS_{regular}$ can be calculated as in Equation 2.7. The concept of the regularly distributed is simple. The final depth of the PR-Qtree must contain enough leaf size to store all points (N). Each increment of depth will increase the leaf size by $\times 4$. So, the final depth (annotate as d') should have enough leaf size; thus should be greater than or equal to $log_4 N$, which base can be converted to two as shown in the equations (for the consistency of using two as the logarithm base throughout this study). Therefore, $FS_{regular}$ should be the immediate integer that is greater than or equal to $log_2 N/2$; or in other words, it should be the ceiling of the fraction (Equation 2.6).

$$4^{d'} \geq N \tag{2.7.1}$$

$$log_4 4^{d'} \geq log_4 N \tag{2.7.2}$$

$$d' \geq log_4 N \tag{2.7.3}$$

$$d' \geq \frac{log_2 N}{log_2 4} \tag{2.7.4}$$

$$FS_{regular} = d' \geq \frac{log_2 N}{2} \tag{2.7.5}$$

doi:10.6342/NTU201802404

While the final scale indicates the times of splitting boxes to contain at most one point in each cell, it gives some idea of how close does the nearest pair of points (or the second nearest pair). This situation also applied in the FS'.

Using Figure 2.15 as example, the FS' of the distribution in (c) is $5/2 = 2.5$. The FS' for the distributions in Figure 2.16a and c are 2.0 ($12/ceil(5.98)$) and 1.8 ($9/ceil(4.32)$), respectively. Through a series of tests on random distribution, the FS' is about 1.8 to 2.0, wheres the cluster distribution have higher FS'.

Critical scaling magnitude

Besides FS, CS will also be influenced by the size of the data points. The higher the number of points, the deeper the CS can be. According to the practical definition, the CS can range from $SS_{observe} + 1$ to $FS_{observe} - 1$. So, CS does not have a smallest or largest possible values as the FS does for a given number of points; which smallest and largest possible values depend on the distribution and is resulted as the $SS_{observe}$ and $FS_{observe}$, respectively. This study follows the concept of the normalization of FS' and defines the critical scaling magnitude (CS') as the comparison of the $CS_{observe}$ to the regularly distributed $CS_{regular}$. The calculation of CS' is shown in Equation 2.8.

$$CS' = \frac{CS_{observed}}{CS_{regula}} \tag{2.8.1}$$

$$CS_{regular} = FS_{regular} - 1 \tag{2.8.2}$$

$$CS_{regular} = ceil(\frac{log_2 N}{2}) - 1 \tag{2.8.3}$$

Because in theory, all neighboring pairs of points in a regular distribution have the same distance, i.e. no pairs are closer than others. Thus, the theoretical $CS_{regular}$ is the largest possible value of CS for the regular distribution, i.e. the $CS_{regular}$ is equal to the $FS_{regular} - 1$. Since the possible range of CS depends on the SS and FS, the CS' range as a floating value that is greater than zero.

The CS' of the distributions in Figure 2.16a and c are 1.2 ($6/5$) and 1.0 ($4/4$), respectively. Through a series of test with random point distribution, the CS' is always

doi:10.6342/NTU201802404

somewhere near 1.0, wheres for clustered distribution, the CS' is higher.

Relative critical scale

As the changes of overall distribution after critical scale (i.e. with smaller boxes) is less significant, this means that with the patterns on the critical scale, the macro pattern can be revealed. On the other hand, the final scale represents the smallest size needed for separating all points into an individual cell. This indicates that the range of scales between critical scale and the final scale shows the range needed for the micro pattern to continue developing. In other words, if the critical scale is nearer to the final scale, the existence of micro pattern is less significant; if the critical scale is nearer to the starting scale, there is a lot of micro patterns to be developed starting from the macro pattern (on CS). Therefore, another additional index on scales is introduced as the relative critical scale (CS^{rel}), i.e. the critical scale relative to the range of starting scale to final scale, which value is range between zero and one. The equation of relative critical scale is shown as follow (Equation 2.9).

$$CS^{rel} = \frac{CS - SS}{FS - SS} \tag{2.9}$$

For the example in Figure 2.16a and c, the CS^{rel} are 0.50 ($6/(12-0)$) and 0.44 (4/(9-0)), respectively.

2.5.3 Point aggregations

The aforementioned concept stated that critical scale shows the macro patterns and increasing the resolution will not increase the macro pattern of the points distribution. This study extends this concept and introduces a set of procedures for aggregating points based on the scales. Thus, the aggregated points in different scales can be used to perform spatial pattern analyses and to be compared with the patterns of original point data.

Based on the PR-Qtree, each depth of the tree indicates a size of the box, i.e. the resolution or scale, and each of the branch at a depth indicates that at least one points located within the corresponding cell. Then, for each occupied box (or branch), the points within the box is aggregated into one representative point, namely the aggregated point,

doi:10.6342/NTU201802404

with the count of points as the weight. This study introduces two ways to generate the locations of representative points, including the grid center approach and mean center approach (Figure 2.17). The grid center approach locates the representative point at the geometry centroid of the grid, which ignores the locations of the points within the box, whereas the mean center approach calculates the mean of x- and y-coordinates as the coordinates of the representative point.

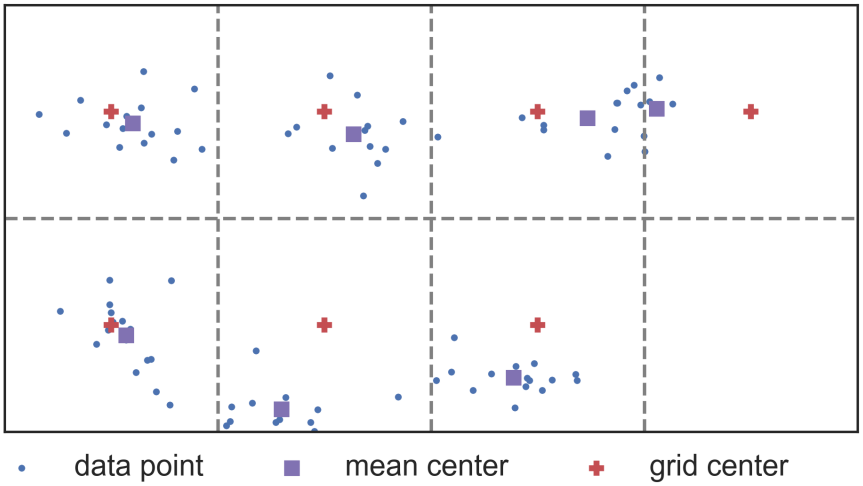


Figure 2.17: This figure demonstrates the two types of center generated from cells based on the points distributions. Grid center is located at the geometry centroid of each cell where there is at least one point fall inside; the mean center is generated based on the mean of the x- and y-coordinates of the points fall inside each cell. No grid or mean center will be generated if there is no points fall inside the cell (e.g. the bottom left cell).

Using John Snow cholera data as an example, Figure 2.18 and Figure 2.19 show the distributions of the aggregated points in different depths using the grid center approach and mean center approach, respectively. On the depth-1 scale, there are four occupied boxes, so four aggregated points were generated. The four points using grid center approach were separated than the distribution using the mean center approach. This is because the original points are more concentrated at the center of the study area, and the mean center approach intended to locate the points at where the original points within each cell are concentrated. This differences between the two approaches applied to depth-2 and depth-3. But, when the depth is increased until higher value, the cells become smaller, that the locations of aggregated points become similar between the two approaches.

Figure 2.20 shows the aggregated points on the critical scale (depth-5) using the two

doi:10.6342/NTU201802404

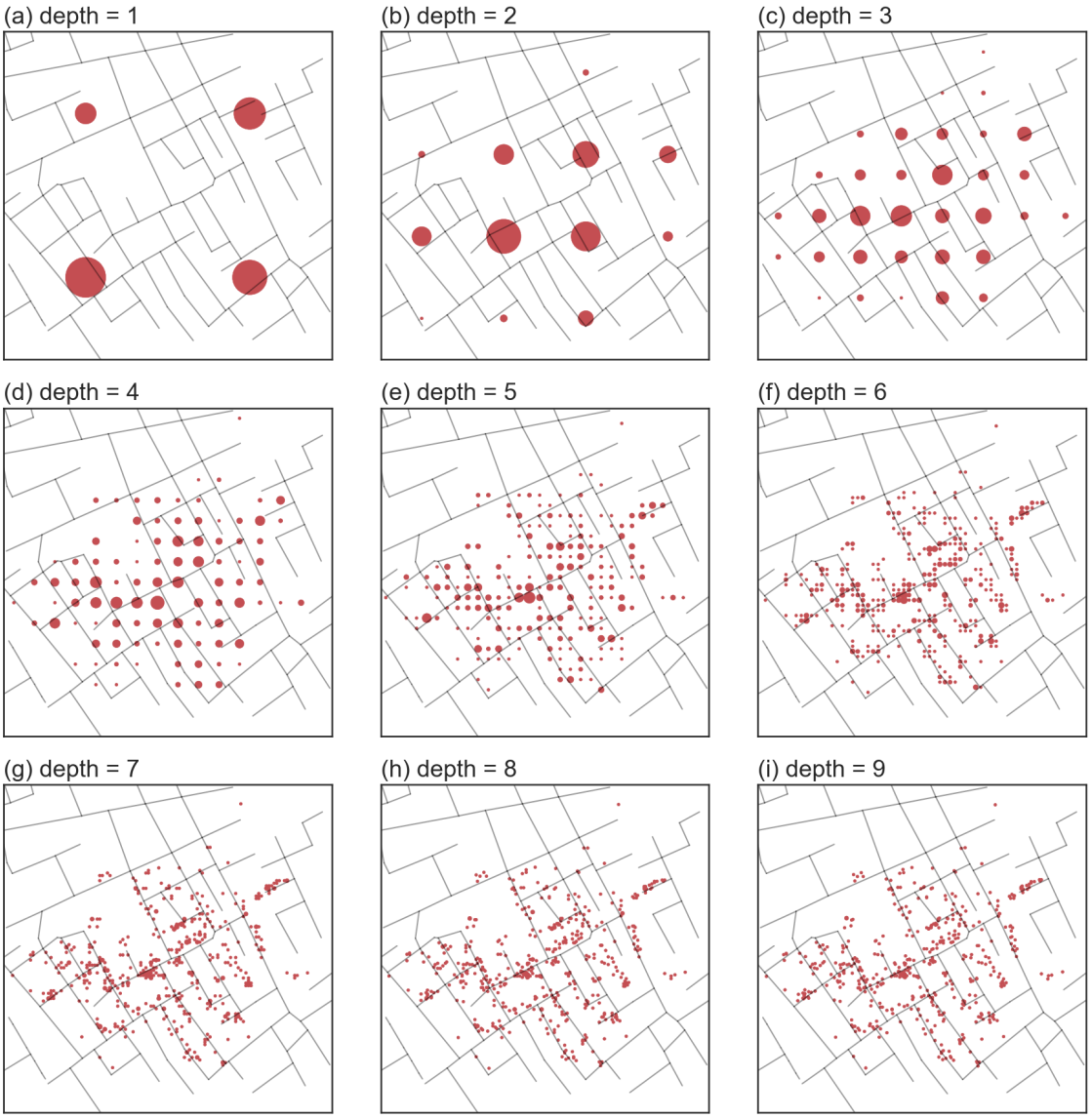


Figure 2.18: The aggregation points (grid center) of the John Snow data in each depth. This figure demonstrates the aggregation points on each depth using the grid center approach. For each depth, the occupied boxes were transformed into the aggregation points, and the number of points inside the boxes is assigned as the weight of the aggregated points.

approaches. On the critical scale, the spatial patterns of the aggregated points are visually similar to the original points, especially the aggregated points that uses the mean center approach.

doi:10.6342/NTU201802404

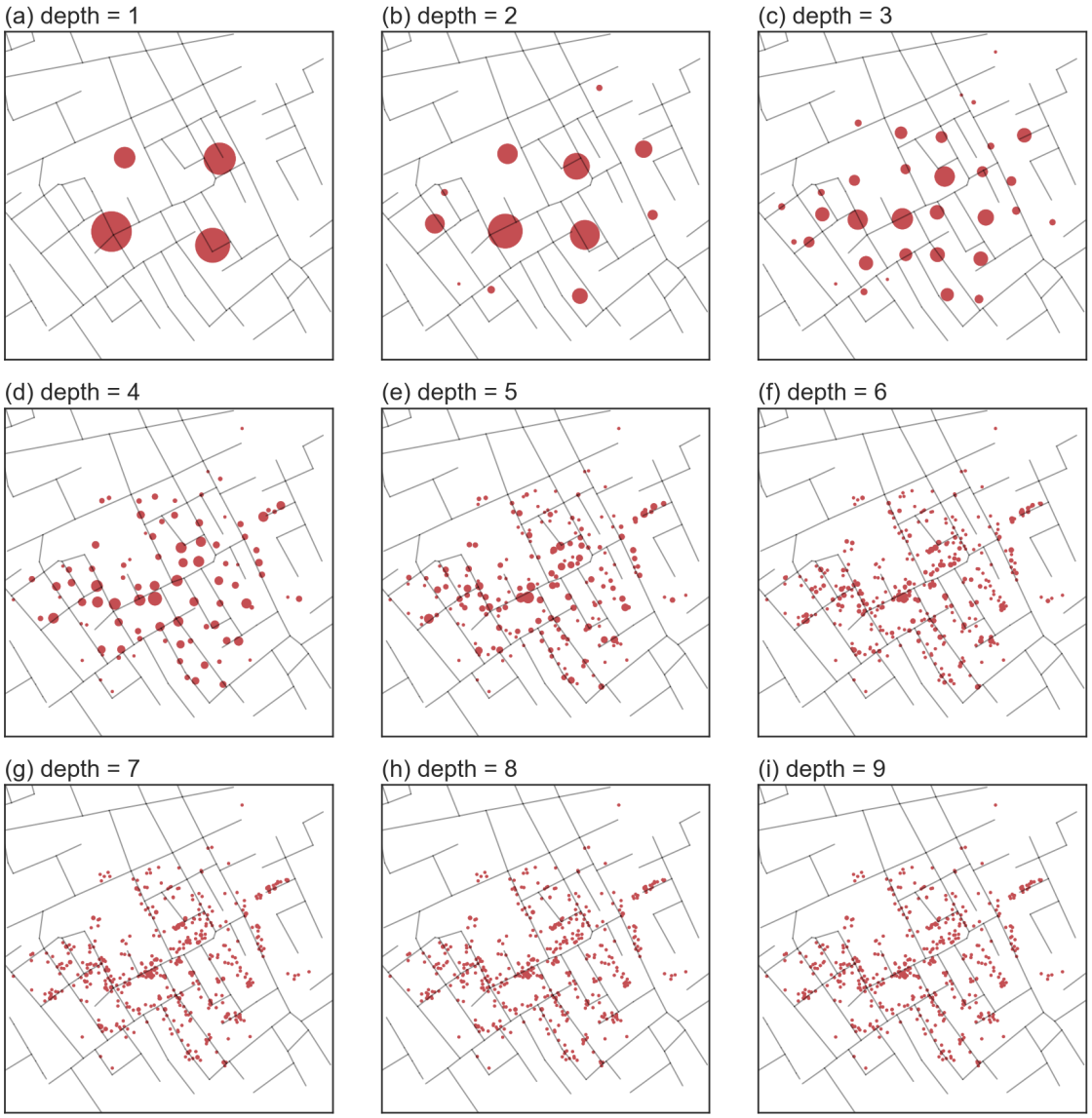


Figure 2.19: The aggregation points (mean center) of the John Snow data in each depth.This figure demonstrates the aggregation points on each depth using the mean center approach.

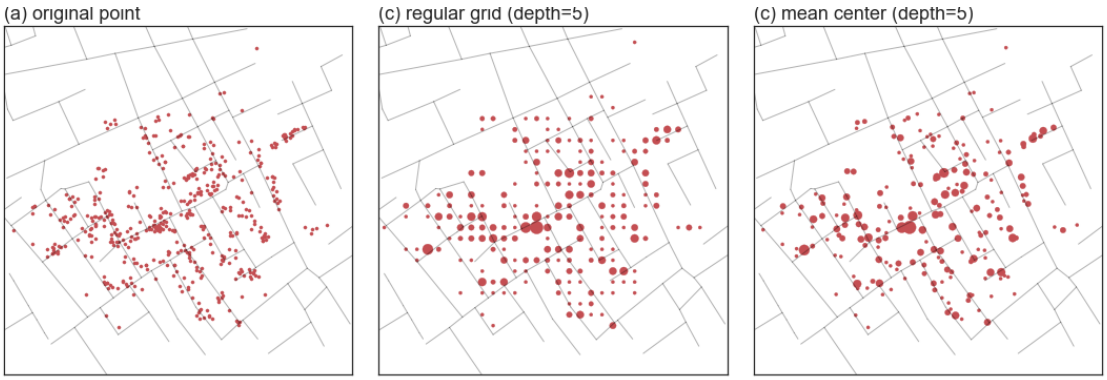


Figure 2.20: The aggregation points on critical depth for the John Snow data. This figure shows: (a) the original points, (b) the aggregated points on critical scale using grid center, and (c) the aggregated points on critical scale using mean center.

doi:10.6342/NTU201802404

Chapter 3

Experiments

Three experiments were designed in this study, including two simulation studies and one empirical study. In the first experiment, three synthetic distributions were simulated to test the aggregation effects, which distribution included dispersed (regular), clustered, and random distribution. In the second experiment, based on a mono-centric clustering model, this study tested the influences of the three clustering properties, including location, covering size, and the number of points, to the scaling results. In the third experiment, this study analyzed three empirical distribution of point patterns, which included locations of post offices, photocopy shops, and beverage shops, to emphasize the usage of the scaling analysis in the real world, and to illustrates the concept of the macro and micro spatial patterns.

doi:10.6342/NTU201802404

3.1 Experiment one: Point aggregation

3.1.1 Aims

The major objective of this experiment is to compare the spatial pattern between the original points and the aggregation points. Each of the depth of PR-Qtree indicates an aggregation spatial scale. Therefore, a series of aggregation points can be generated based on the structure of PR-Qtree for each scale (depth). Each of the aggregation points set shows a spatial pattern. With higher scale (smaller cell size), the resulting spatial pattern will be more similar to the spatial pattern of the original points. And thus, the research question of this experiment is that: how do the differences between the aggregation points and original points changes against the increment of scales, and how does the performance of the aggregation points on critical scale shows.

As discussed in previous sections, the main idea of the critical scale is the spatial scale that can capture the macro pattern of point distribution. In other words, the aggregation points using the critical scale should show a similar spatial pattern to the original points (input data) does. To evaluate this idea and the capability of critical scale in terms of capturing the macro pattern, this experiment is designed to used several point pattern analyses, including three global pattern analyses and two micro pattern analyses, to analyze the spatial pattern of aggregation points, and compare these result with the spatial patterns calculated based on original points.

3.1.2 The three cases

The cases in this experiment included three theoretical point distribution. The distribution is shown in Figure 3.1. The study area is designed to have 1000×1000 spatial units and contains 400 points. The three cases have different spatial pattern settings, including dispersed (regularly arranged), clustered, and random. The first case (disperse) shows a regularly gridded points, that all points are divided into 20 rows and 20 columns, and each pair of the horizontally and vertically neighboring points have the same distance. The

[1]The algorithm for generating clustered distribution is shown in Appendix I. The second case uses the parameter set ($N = 400, C(500, 500), P_{sigma} = 0.125, Box(0, 0, 1000, 1000)$).

doi:10.6342/NTU201802404

second case (clustered) is concentrated at the center of the study area (coordinates (500, 500)), that have higher density at the centroid of cluster than the outer ring.[1] The third case (random) shows a random distribution that is generated based on a uniformly distributed probability function. In other words, each point would have the same probability of appearing in every part of the study area.

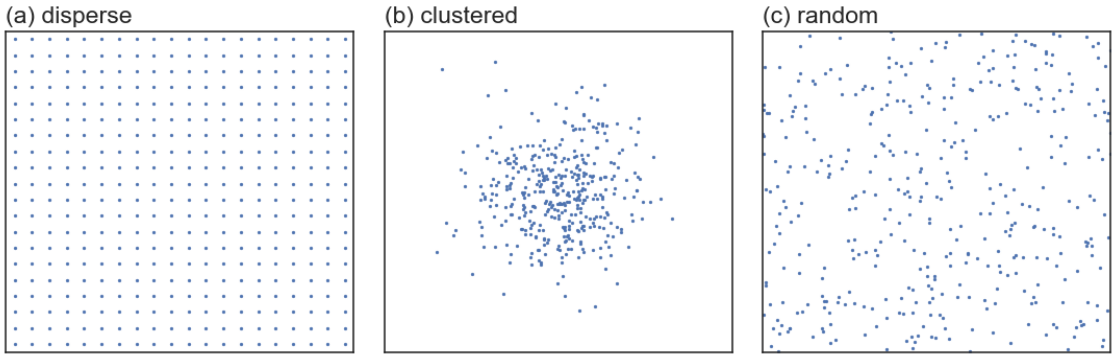


Figure 3.1: The point distribution of the three cases: (a) disperse case, (b) clustered case, and (c) random case. The three cases have the same size of study area, and the same number of points.

3.1.3 Experiment design

In this experiment, the three cases were used to run several spatial pattern analyses, which included global and local analyses. The global analysis methods included the K-order nearest neighbor analysis (NNA), K-function, and Weighted K-function. The local analysis method included the kernel density estimation and weighted kernel density estimation. The concepts of the five analyses were briefly introduced and the practical application in this experiment was described in the following.

The nearest neighbor analysis is a framework that calculates the average distance for each point to its nearest neighbor (Clark and Evans, 1954; Boots and Getis, 1988; Gatrell et al., 1996). This analysis framework is designed to compare this average distance to the expected average distance based on random distribution. Thus, the basic nearest neighbor analysis will continue the calculation and produce a nearest neighbor index, which is an indicator to evaluate if the point distribution is globally clustered. K-order nearest neighbor analysis, or high order nearest neighbor analysis (i.e. using K to indicate the level of

doi:10.6342/NTU201802404

order), is an extended version of nearest neighbor analysis (Wong and Lee, 2005). The first order (K=1) nearest neighbor index indicates the basic nearest neighbor index. The second order (K=2) will use the distances to the second nearest neighbor for the calculation instead of the first nearest neighbor, and the third order (K=3) will use the third nearest distance, etc. Therefore, K-order nearest neighbor analysis will produce a line that shows the nearest neighbor index against the K-order as the main result. In this study, the focus is not on whether the distribution is a cluster. Therefore, the average distance of each point to the K-order nearest neighbor is used as the NNA findings in this experiment.

K-function, or Ripley's K-function, is also known as the second-order analysis of point pattern (Getis and Franklin, 1987). Instead of measuring the distance between points, the concept of K-function count the number of points falls inside a search radius from each point, i.e. the number of neighbors within a spatial lag area for each point; then, plots the average number of neighbors against the search radius (Wong and Lee, 2005). Therefore, K-function is designed to understand on average, how many neighbors a point has for a series of given spatial-lags. If a point has a lot of neighbors within a close distance, then the distribution has a global clustering pattern on the distance. While K-function counts the number of neighbors, the weight of point should be considered in some cases. Therefore, a weighted version of K-function (Getis, 1984) is introduced. Instead of counting the number of neighbors, the Weighted K-function quantifies the interactions between two weighted points by multiplying the weights of each point that falls within a spatial lag from a point to the weight of the target point. Both of the results of K-function and weighted K-function are converted into L functions for a better understanding and visualization.

The above three spatial pattern analyses focus on the global pattern of a point distribution. Kernel density estimation (KDE) is a method for estimating the kernel density of points for each part of the study area, i.e. to reveal the local pattern (Bailey and Gatrell, 1995; Seaman and Powell, 1996; Anderson, 2009; Gerber, 2014). Given a set of points that falls within a study area, the KDE is designed to estimate the two-dimensions probability density function throughout the study area. Then, the density values can be calculated for each cell of a predefined regular lattice, which size should be small enough for generating

doi:10.6342/NTU201802404

a smooth and continuous surface of density. The weighted version of KDE includes the differences of weights in the calculation.

3.1.4 Results

The calculation results of the scaling analysis framework for the three cases are shown in Table 3.1. Figure 3.2 shows the OB-plot and OR-plot of the three cases. The final scales of the three cases were five (dispersed, $FS' = 1.0$), nine (clustered, $FS' = 1.8$), and nine (random, $FS' = 1.8$). Given the same number of nodes, the final scale was influenced by the nearest (or second nearest) pair of points. Since the disperse case had all points evenly spaced, the PR-Qtree resulted in a shallower structure (Figure 3.2a and d), which indicated a smaller value of depth (final scale). The final scales of clustered and random cases were same, which was because the clustered cases did not contain an extreme close distribution; the random distribution, on the other hand, would have some points closer to each other and some further to each other (Figure 3.2c and f).

Focusing on the critical scale (and critical scaling magnitude), the three cases had similar results, which is 4 (1.0), 5 (1.25), and 4 (1.0), which indicated that the critical scale for capturing the macro pattern was similar through the three cases. Based on the critical scaling magnitude, the CS' results of the dispersed and random cases showed that they are same with the uniformly distributed patterns. This might because the random case was designed to use the uniformly distributed probability function, i.e. all locations had the same probability of having a point. The relative critical scale, on the other hand, showed a high value for the dispersed case (0.8), and a moderate value for the clustered and random cases. This indicated that most of the scales for disperse case had the similar spatial pattern in terms of the big picture of the distribution.

Using the points aggregation methods as mentioned in the previous section, the aggregation points for the depths before reaching the final scale were generated for the three cases. Figure 3.3 shows the aggregation points in the four scales. The final scale (depth-5) was skipped because the final scale indicated that all points are separated into a cell, i.e. the aggregation points if generated, was the same as the input points (under mean center

doi:10.6342/NTU201802404

Table 3.1: The number of points in each cases and category.

index	disperse	clustered	random
no. node	400	400	400
Starting Scale	0	0	0
Critical Scale	4	5	4
Final Scale	5	9	9
Relative CS	0.8	0.56	0.44
FS magnitude	1.0	1.8	1.8
CS magnitude	1.0	1.25	1.0
OB slope	2.0	1.51	1.93
OB r2	1.0	0.992	0.999
OR slope	1.36	1.77	1.84
OR r2	1.0	0.995	0.995

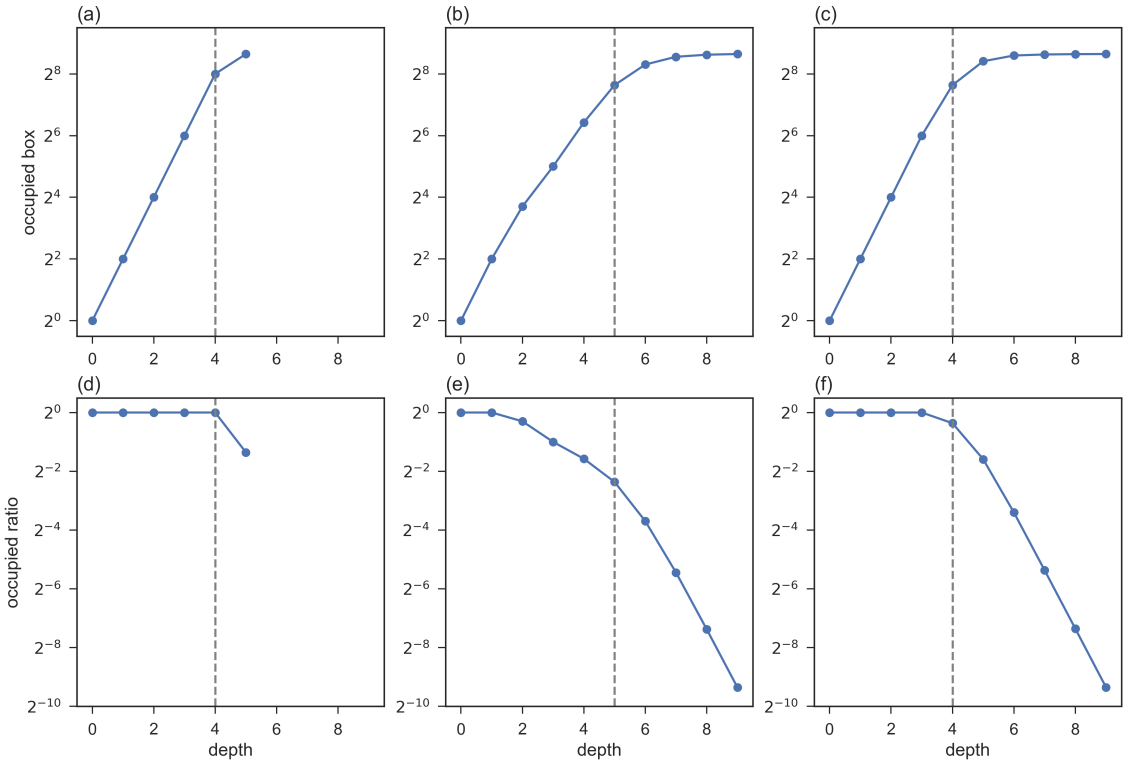


Figure 3.2: The OB-plot and OR-plot of the three cases, including the critical scale (dashed lines). Case 1 (a and d) stopped at a shallower depth, with the critical scales near the final scale. Case 2 (b and e) shows a deeper depth, and the critical scale is also relatively deeper. Case 3 (c and f) has a clear pattern than case 2, which turning points on both figure OB and OR plot are clearer, and is a little shallower than case 2.

approach). The two rows in the figure indicated the aggregation using grid center approach (first row) and mean center approach (second row). All of the aggregated points in the first case were also evenly spaced, i.e. regularly distributed in all boxes on each depth. There were no visual differences between the grid center approach and mean center approach.

doi:10.6342/NTU201802404

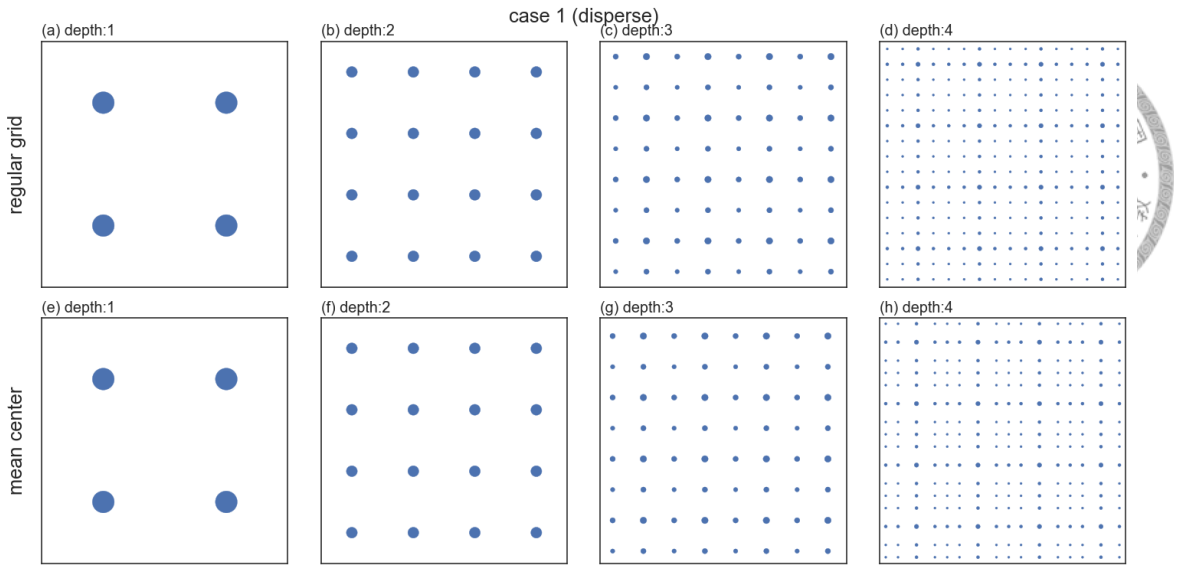


Figure 3.3: The aggregation points of case 1 in experiment 1. The aggregation points in the first row used grid center approach, whereas the second row used mean center approach.

The final scale of the second case was nine. Therefore the aggregation points for the first eight depths were generated (Figure 3.4). Since the original points were set to be concentrated at the center of the study area, the aggregation points were also tended to be concentrated at the center, with significantly higher weights on the points appear in the four boxes at the middle of the study area on the first two depths. As the depth increase, the weights started to be divided into smaller pieces of aggregation points because of the splitting of boxes. On deeper depth, the hierarchical of the weight concentrations became clearer, which forms a radiation shapes that had higher weight at the center, and decreased with the distance to the center. This pattern showed consistency with the main concept of the random clustering model setting (as described in Appendix I). The differences between the grid center and mean center become smaller as the depth is increasing.

The final scales of the third case was also nine, and eight of the scales were used to generate aggregation points (Figure 3.5). The aggregation points of the third case were similar to the first case in terms of evenly distributed but with some absence of points at some random locations. This might because the distribution of random was built on top of the uniformly distributed probability. When the distribution is viewed from global scale to local scale, the pattern of random occurs by some the disappearance of points at some locations, which may move to somewhere close to a randomly selected point, and the re-

doi:10.6342/NTU201802404

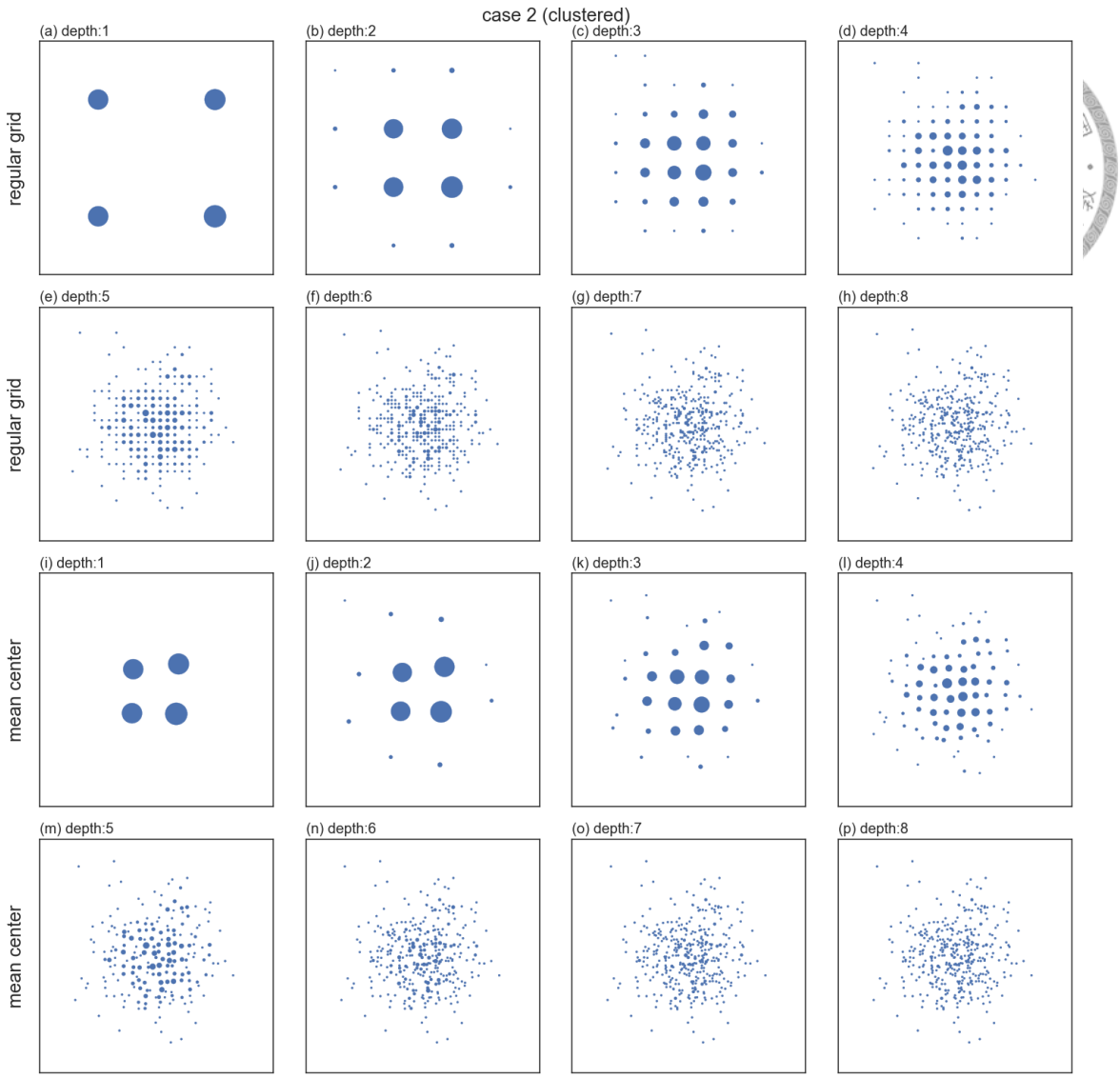


Figure 3.4: The aggregation points of case 2 in experiment 1. The aggregation points in the first two rows used grid center approach, whereas the third and fourth rows used mean center approach.

sulting pattern is a micro pattern that is not recognizable when the scale is lower (shallower depth). By comparing the mean center approach (MC) to grid center approach (GC), the aggregation points using MC looked a little oscillated from the grid center locations.

doi:10.6342/NTU201802404

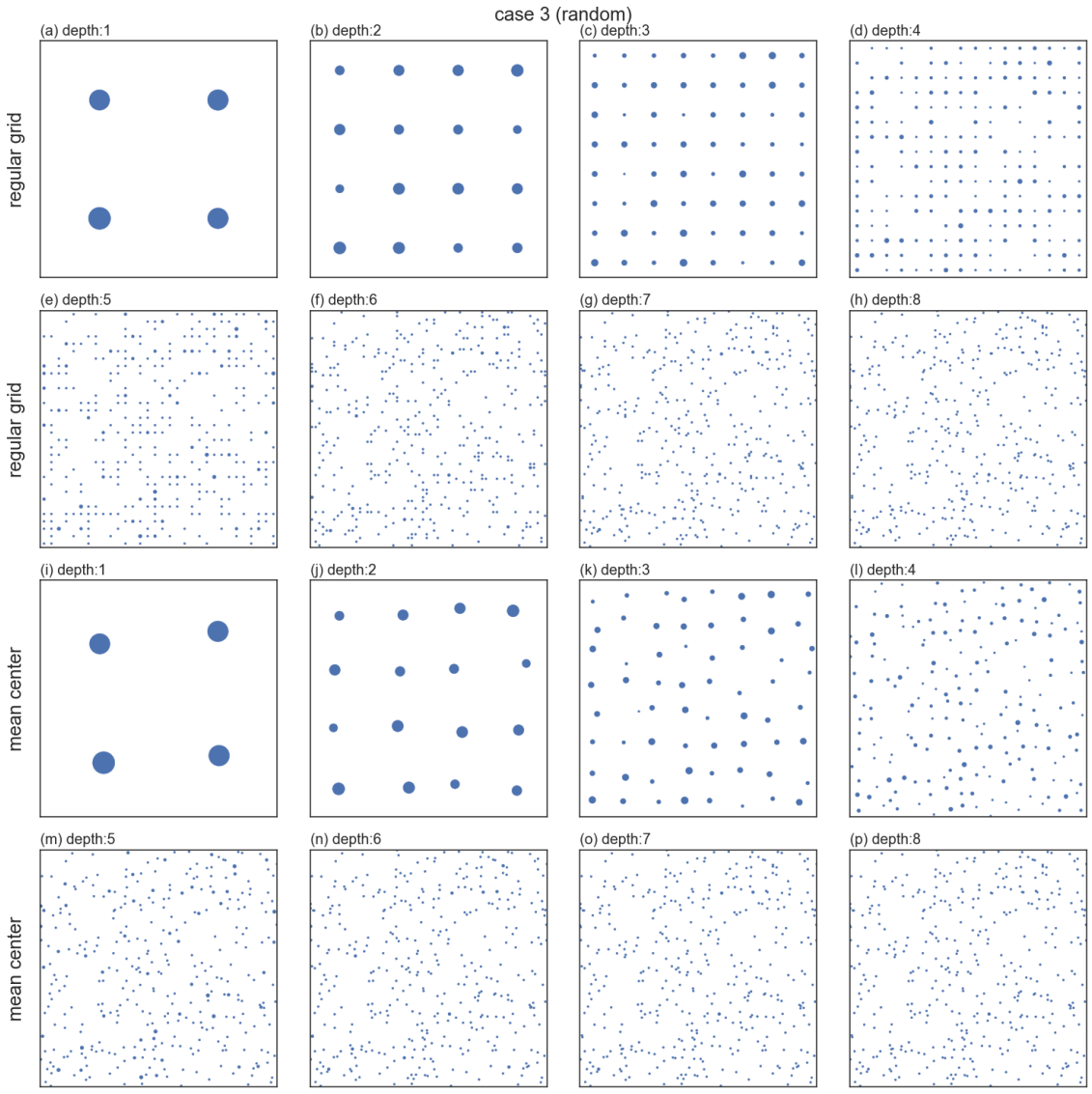


Figure 3.5: The aggregation points of case 3 in experiment 1. The aggregation points in the first two rows used grid center approach, whereas the third and fourth rows used mean center approach.

doi:10.6342/NTU201802404

The processes of the five sets of spatial pattern analyses using the series of aggregation points were conducted. The intermediate results were shown in Appendix II, including the aggregation of points using grid center and mean center approaches. The results of the aggregation points were compared to the spatial pattern results using the original points. The normalized root-mean-square-errors (RMSE) were shown in Figure 3.6.

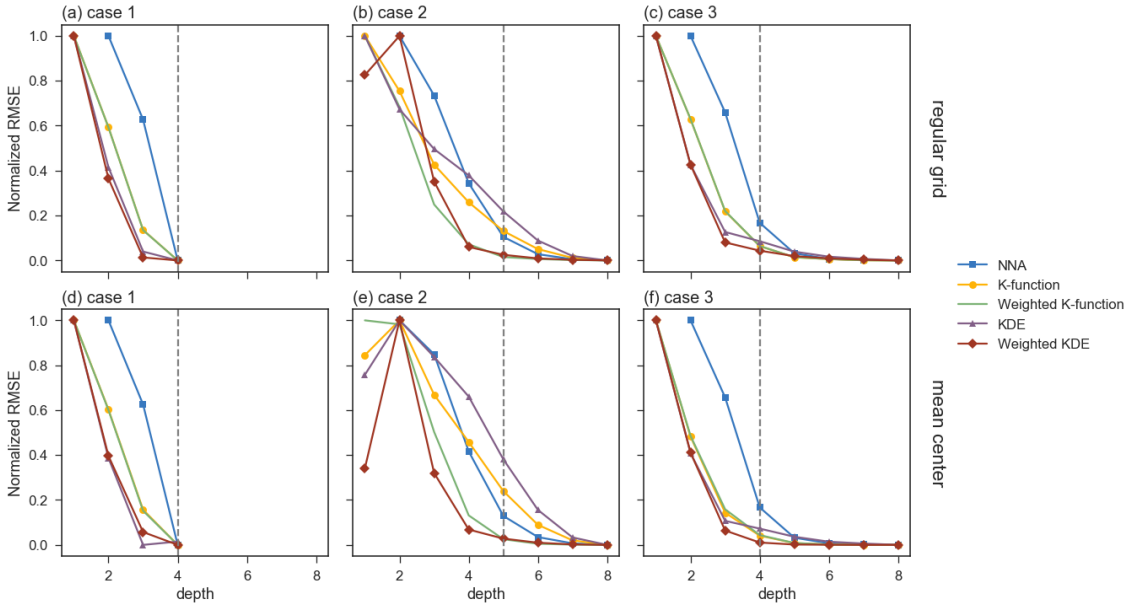


Figure 3.6: The Normalized RMSE for the five analyses on the three cases in experiment 1. The results in the first row is based on grid center approach, whereas the second row is based on mean center approach.

The RMSE results of all three cases suggested that the differences between the spatial patterns of the original distribution and the aggregated points dropped quickly with steeper slope before the critical scale, and reached a low level of differences on the critical scale. For the second case and third case, the RMSE kept decreasing after the critical scale with a gentle slope; for the first case, the PR-Qtree reached the final scale after the critical scale. Among the five analyses, the two weighted analyses, including weighted K-function and weighted KDE, had a steeper dropping slope and reached a low level of differences faster than their unweighted counterparts; NNA had the slowest decreasing speed. The decreasing trends against scales (depths) was clearer for the first and third cases. For the second case, the lines of the decreasing RMSE did not join together as they did for the other two cases. This situation indicated that the spatial patterns between the five analyses were different, and which might need one scale deeper for a better capturing of the

doi:10.6342/NTU201802404

original spatial pattern for some of the analyses (e.g. the unweighted version of K-function and KDE). The reason behind the situation that weighted version of K-function and KDE dropped faster might because of the number of points that merged into the aggregated points were converted into a weight and is considered in the point pattern analysis; whereas the unweighted version of analyses ignored the number of points and focused only on the locations.

Two aggregation approaches had a slight difference in RMSE, while their average changing trends against depth were similar. Although the mean center approach was expected to have better results for capturing the original pattern, the results suggested that the grid center approach had similar results on higher scales, e.g. critical scale (or higher).

Figure 3.7 shows the original points and the aggregation points on the critical scale using the two approaches (GC and MC). The locations of the grid centers were slightly different from the mean centers, which was more similar to the original points. But, the results suggested that aggregation using GC on the critical scale is enough in terms of capturing the original distribution.

doi:10.6342/NTU201802404

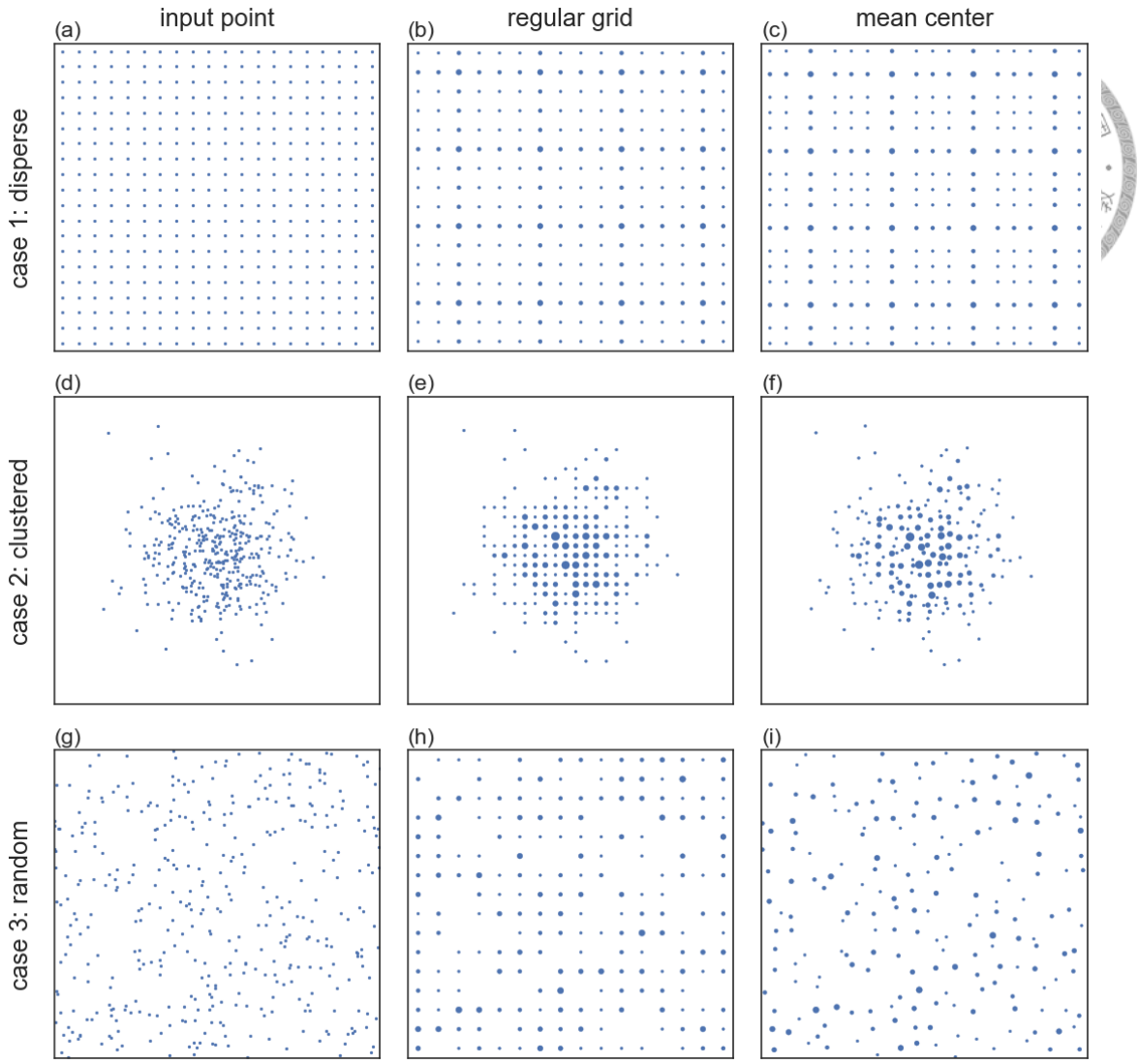


Figure 3.7: The comparison of aggregation points on critical scales in experiment 1. The original points were shown in the first column, the aggregation points on the critical scale using grid center approach and mean center approach are shown in second and third columns. The three cases were separated in three rows.

3.1.5 Summary

The scaling analysis framework using the PR-Qtree structure was tested on three theoretical distribution in the first experiment. The major aim of this experiment was to test the aggregation effect on the scaling process. The summary of the findings is listed as below:

1. The FS of the random case was about two times the FS of the dispersed case ($FS' = 1.8$ vs. $FS' = 1.0$), while their CS was about the same. In other words, the macro pattern of the random case could be described using the same scale as the dispersed case does; the difference between the random case and disperse case was on the micro pattern, i.e. the

doi:10.6342/NTU201802404

tail parts after CS in the OR-plot.

2. The analyses of aggregation effect showed that the spatial patterns of the original points could be captured with the aggregation points on the critical scale. The weighted version of analyses on aggregation points had better results, which merged the close points (locally clustered) into an aggregated point with the weight as the count of merged points. In other words, the low RMSE indicated that the locations of these close points under the critical scale do not have the ability to influence the macro spatial patterns; they can be merged when the study aims are not focused on the local anomalies pattern.

3. In the case 2, the different analyses resulted in different decreasing trends of RMSE, and the critical scale located somewhere middle of the decreased RMSE. This suggested that the critical scale found from the scaling framework in this study should not be treated as the only one scale; a range of critical scale ± 1 was a reasonable scales for generating aggregation points for different types of analysis.

4. The aggregation points using both GC and MC approaches resulted in a sufficient performance in terms of capturing the macro patterns. The MC approach was expected to have better performance than the GC approach because the representation points of MC considers the locations of points within the cell. But, the GC approach, which has simplified the location of representation points as the geometric centroid of the cells, was also resulted in high performance on the critical scale. This might because of the corresponding cell size for the critical scale is small, which leads to the situation that the location of grid center is not far from the mean center while viewing from the overall perspective.

doi:10.6342/NTU201802404

3.2 Experiment two: Clustering properties

3.2.1 Aims

The major objective of this experiment is to evaluate the influences of clustering properties on the critical scale and other scaling results. A cluster can be described in three perspectives, including the location of a cluster, the area it covers (the area size), and how many points are in the cluster. The clusters at varying locations may be experiencing different levels of the effects from the shape and locations. The cluster that is larger in area size may have less density than the cluster with the same number of points but with a smaller cover area, and thus their critical scale may be different. Given a fixed area, a cluster with more points should have more points near to each other compared to a cluster with fewer points, which may lead to the higher critical scale for viewing the macro pattern. The critical scale is expected to change with different clustering properties, but the effects and changing rates of the clustering properties are unknown.

The knowledge of knowing these effects and changing rates of the clustering properties to the critical scale and other scaling properties can provide more insights on the understanding of the scaling of point patterns. Therefore, this experiment is designed to test and evaluates the changes of critical scales (and other scaling indexes) against the three dimensions of clustering properties.

3.2.2 Experiment design

Based on the three dimensions, including the location, area size, and the number of points, this experiment contains four parts, each of these focused on one dimension of the clustering properties:

- to compare the effects of varying cluster location on the scaling phenomenon;
- to evaluate the influences of cluster area sizes on the scaling phenomenon;
- to test the effects of the number of points in clusters on the scaling phenomenon;

doi:10.6342/NTU201802404

- to test the scaling phenomenon in random distributions with a varying number of points.

The first three parts focus on the three clustering properties, while the last part of the experiment is an extended experiment of the third part, which is also focusing on the number of points but with a different spatial organization. While analyzing each part, the other properties are set to be fixed, and the targeted dimension is set to be changed at a rate.

After the scaling analysis is done, the six scaling results are plotted for observing the changes against the targeted dimension of each part, including the starting scale (SS), critical scale (CS), final scale (FS), relative critical scale (CS rel.), critical scaling magnitude (CS mag.), and final scaling magnitude (FS mag.).

3.2.3 The cases

The clusters of each part is generated using the same cluster generation model as the second case in the previous experiment, which is described in Appendix I. The cluster generation model needs three main parameters, which is the number of points (N), the coordinates of the cluster center ($C(x, y)$), and the degree of separation (P_{sigma}). These three parameters are used to generate the clusters for each part of the experiments. The other parameter is the setting of the study area, which is fixed at the size of 1000×1000. Although each of the following parts focuses only on one of the dimension, the point generation using each of the parameters combinations ($5 \times 7 \times 8$) are all generated, and the results are shown in Appendix III.

In the first part, the focus was on the changing of location, therefore the parameter of the center (C(x,y)) was changed to generate clusters located at different locations, whereas the other two parameters were fixed ($N = 512$, $P_{sigma} = 2^{-4} = 0.0625$). The center was set to be five coordinates, that were started from the corner towards the center. The center coordinates are shown in Table 3.2. Each of the cluster settings was used to generate 99 point sets, which was used to run the scaling analysis framework, and the results were converted to means (of the 99 sets) for showing the average pattern along the centers.

doi:10.6342/NTU201802404

A set of point distribution using the five centers and the fixed parameters are shown in Figure 3.8.

Table 3.2: The parameter settings for part 1 experiment.

	case id	coordinate	P_{sigma}	no. points	note
1	31.25	(31.25,31.25)	$2^{-4} = 0.0625$	512	at the corner
2	62.5	(62.5,62.5)	$2^{-4} = 0.0625$	512	
3	125.0	(125.0,125.0)	$2^{-4} = 0.0625$	512	
4	250.0	(250.0,250.0)	$2^{-4} = 0.0625$	512	
5	500.0	(500.0,500.0)	$2^{-4} = 0.0625$	512	center of study area

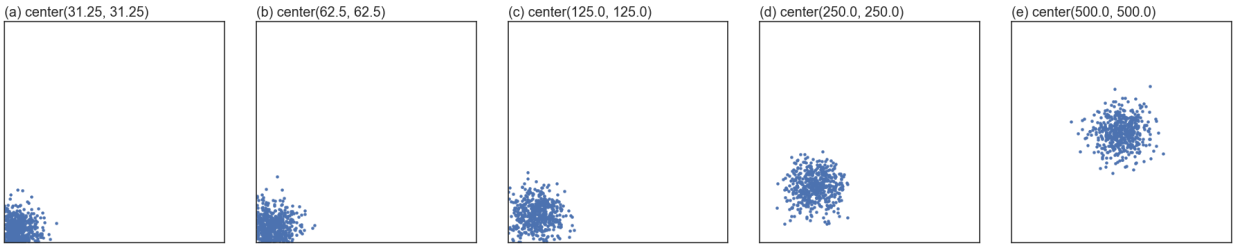


Figure 3.8: A demonstration of generated points for part 1 in experiment 2. The location of center is at the corner (a), and moving forward to the center of the study area (e). The other parameter settings are fixed.

For the second part, the aim was to test the influence of the area size of the cluster on the scaling analysis. The changes in area size could be captured by changing the degree of separation parameter in the model. In this part of the experiment, the number of points was fixed at 512 and the location of the cluster center was set at c(500,500). The degree of separation was set to a range from the smallest 0.015625 to the largest 1.0, with a total of seven steps. The parameter settings is shown in Table 3.3. Each of the parameter settings was repeated for 99 times for calculating the mean and standard deviation. A set of the demonstration for the part 2 experiment is shown in Figure 3.9.

doi:10.6342/NTU201802404

Table 3.3: The parameter settings for part 2 experiment.

	case id	coordinate	P_{sigma}	no. points	note
1	2^{-6}	(500.0,500.0)	$2^{-6} = 0.015625$	512	the smallest size
2	2^{-5}	(500.0,500.0)	$2^{-5} = 0.03125$	512	
3	2^{-4}	(500.0,500.0)	$2^{-4} = 0.0625$	512	
4	2^{-3}	(500.0,500.0)	$2^{-3} = 0.125$	512	
5	2^{-2}	(500.0,500.0)	$2^{-2} = 0.25$	512	
6	2^{-1}	(500.0,500.0)	$2^{-1} = 0.5$	512	
7	2^{0}	(500.0,500.0)	$2^{0} = 1.0$	512	the largest size

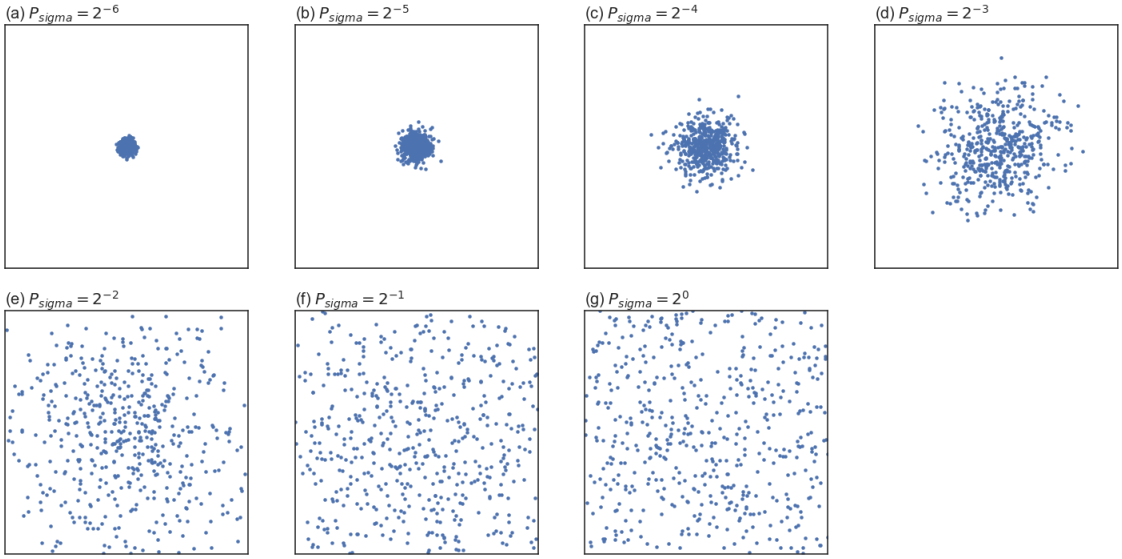


Figure 3.9: A demonstration of generated points for part 2 in experiment 2. The cover area of the cluster is set to be small with lower degree of separation (a), and becoming larger until it is larger than the study area and forms a random distribution (e). The other parameter settings are fixed.

The third part of the experiment two was to test the effect of a different number of points in the study area to the scaling analysis result. Changing the number of points within the fixed area size of the cluster means changing the density of the whole point distribution. In the third part, the number of points was set to be the lowest from 16 points to the highest 2048 points and a total of 8 steps. The other parameters were fixed (center of the cluster was C(500,500), and the degree of separation $P_{sigma} = 0.0625$). The full parameter settings for the third part is shown in Table 3.4. The point distribution of the generated points is demonstrated in Figure 3.10.

doi:10.6342/NTU201802404

Table 3.4: The parameter settings for part 3 experiment.

	case id	coordinate	P_{sigma}	no. points	note
1	2^4	(500.0,500.0)	$2^{-4} = 0.0625$	$2^4 = 16$	the lowest
2	2^5	(500.0,500.0)	$2^{-4} = 0.0625$	$2^5 = 32$	
3	2^6	(500.0,500.0)	$2^{-4} = 0.0625$	$2^6 = 64$	
4	2^7	(500.0,500.0)	$2^{-4} = 0.0625$	$2^7 = 128$	
5	2^8	(500.0,500.0)	$2^{-4} = 0.0625$	$2^8 = 256$	
6	2^9	(500.0,500.0)	$2^{-4} = 0.0625$	$2^9 = 512$	
7	2^{10}	(500.0,500.0)	$2^{-4} = 0.0625$	$2^{10} = 1024$	
8	2^{11}	(500.0,500.0)	$2^{-4} = 0.0625$	$2^{11} = 2048$	the highest

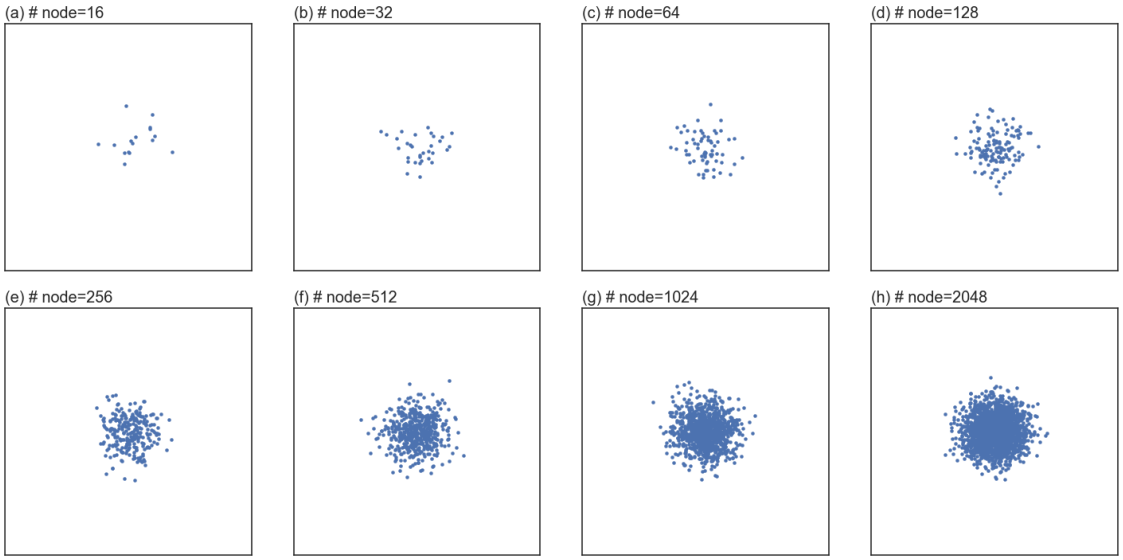


Figure 3.10: A demonstration of generated points for part 3 in experiment 2. The number of points is started as 16 as the lowest (a) to 2048 as the highest (f). The different number of points in part 3 form a same area size cluster within the fixed study area.

Part four was an extended analysis of the third part, that the distribution was set to have a significantly large separation degree, which was used to demonstrate the random distribution. While the aim of the fourth part was also to test the effect of changing the number of points to the scaling analysis results, it was also to compare the effects between the third and the fourth parts. The number of points and other parameter settings was set to be same as the third part, except for the degree of separation (P_{sigma}=1.0). The detail of the setup is shown in Table 3.5. The demonstrations of the generated points are shown in Figure 3.11.

doi:10.6342/NTU201802404

Table 3.5: The parameter settings for part 4 experiment.

	case id	coordinate	P_{sigma}	no. points	note
1	2^4	(500.0,500.0)	$2^0 = 1.0$	$2^4 = 16$	the lowest
2	2^5	(500.0,500.0)	$2^0 = 1.0$	$2^5 = 32$	
3	2^6	(500.0,500.0)	$2^0 = 1.0$	$2^6 = 64$	
4	2^7	(500.0,500.0)	$2^0 = 1.0$	$2^7 = 128$	
5	2^8	(500.0,500.0)	$2^0 = 1.0$	$2^8 = 256$	
6	2^9	(500.0,500.0)	$2^0 = 1.0$	$2^9 = 512$	
7	2^{10}	(500.0,500.0)	$2^0 = 1.0$	$2^{10} = 1024$	
8	2^{11}	(500.0,500.0)	$2^0 = 1.0$	$2^{11} = 2048$	the highest

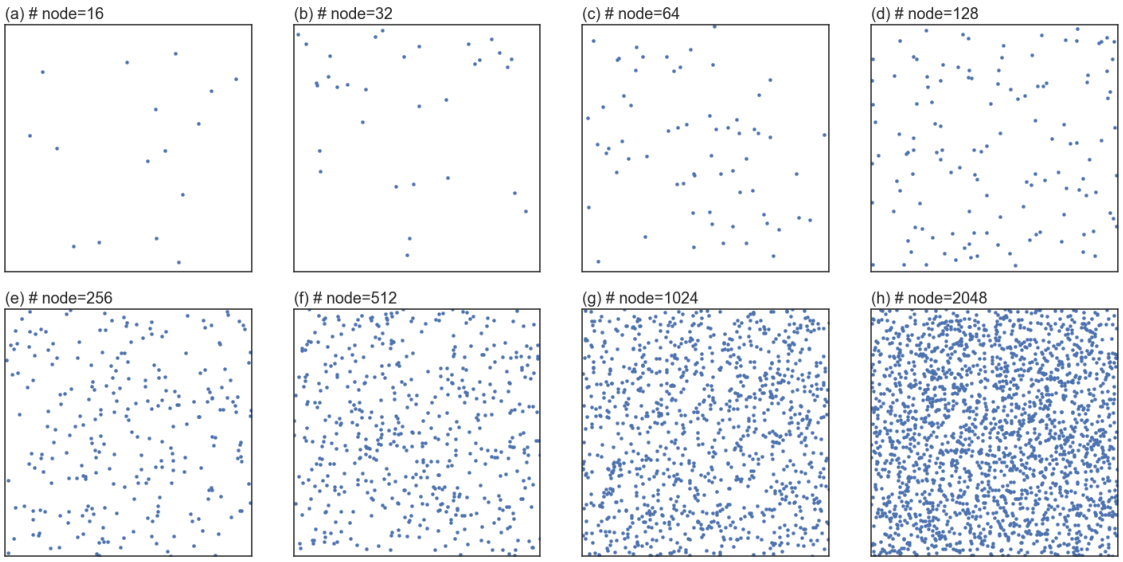


Figure 3.11: A demonstration of generated points for part 4 in experiment 2. The number of points is started as 16 as the lowest (a) to 2048 as the highest (f). The different number of points in part 4 form a random distribution within the study area.

3.2.4 Results

Figure 3.12 shows the effect of changing cluster location to the scaling analysis results, which details is shown in Table 3.6. While the cluster was at the corner, critical scale and final scale were relatively higher than when it was at the center. This situation might cause by the result that the cluster area was smaller, that the cluster shape was similar and the number of points is the same, the truncation of the area leads to the higher density for the cluster at corner. The starting scale of the corner cluster was higher due to the possibility of having the other three cell on first depth were empty in some of the point generations of the 99 times, which pull up the mean of starting scale. The differences in three scales against the changing of the center were on average not higher than one depth.

doi:10.6342/NTU201802404

This suggested that the effects of location on the scales were less significant.

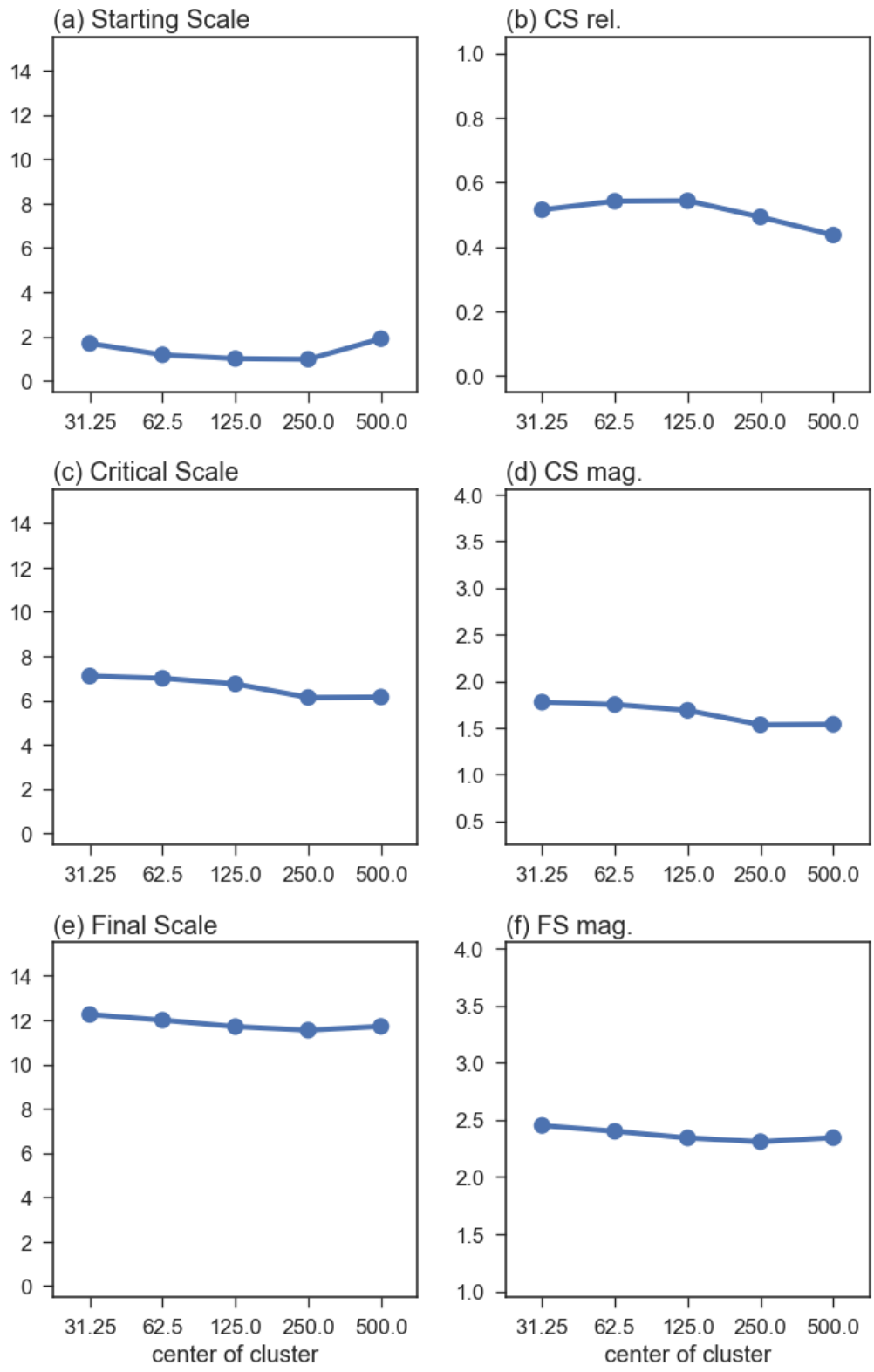


Figure 3.12: The scaling results of part 1 in experiment 2. The changes of the six indexes were shown against the changes of cluster location, which is starting from corner ($C(31.25, 31.25)$) to center ($C(500, 500)$).

doi:10.6342/NTU201802404

The extended scaling results, including the relative critical scale, critical scaling magnitude, and final scaling magnitude, provided some information on the distribution. The relative critical scale was about 0.4 to 0.6 while the cluster shape was set as a moderate size cluster. The critical scaling magnitude was about 1.5 at corner and keeps dropping until 1.3 at the center. The final scaling magnitude was also dropping from about 2.3 (corner) to 2.1 (center).

Table 3.6: The details of the scaling results for part 1 in experiment 2.

center	SS	CS	FS	CS rel.	CS mag.	FS mag.
31.25	1.69(±0.47)	7.1(±0.3)	12.25(±0.91)	0.51(±0.06)	1.78(±0.08)	2.45(±0.18)
62.5	1.17(±0.38)	7.0(±0.0)	12.0(±0.95)	0.54(±0.05)	1.75(±0.0)	2.4(±0.19)
125.0	1.0(±0.0)	6.75(±0.44)	11.71(±1.0)	0.54(±0.07)	1.69(±0.11)	2.34(±0.2)
250.0	0.97(±0.17)	6.13(±0.34)	11.55(±0.99)	0.49(±0.06)	1.53(±0.08)	2.31(±0.2)
500.0	1.9(±0.44)	6.15(±0.36)	11.72(±1.06)	0.44(±0.07)	1.54(±0.09)	2.34(±0.21)

The second part results showed a significant effect of the changing cluster area size (Figure 3.13, Table 3.7). When the cluster was small (in terms of area), the distribution was concentrated at some of the cells at the center, leaving the cells at the outer ring empty. This distribution caused the starting scale of the cases with smaller area size ($P_{sigma} \leq 2^{-4}$) were higher ($0 < SS \leq 4$). The critical scale and final scales started decreasing (from $CS = 8$ and $FS = 14$) when the size was the smallest until when the degree of separation $P_{sigma} = 2^{-2}$ (to $CS = 4$ and $FS = 10$), and the effect disappear where $P_{sigma} > 2^{-2}$ (i.e. $CS = 4$ and $FS = 10$). This suggested that the critical and final scales would change sensitively with the changing area size. While the increment of the degree of separation (P_{sigma}) was designed as a log_2 steps, the linear result indicated that the effect on CS and FS was a function of the log_2 of the CS or FS. When $P_{sigma} > 2^{-2}$, the distribution was visually showing a random distribution (Figure 3.9), and this might be the reason that the effect of area size to the CS and FS disappeared on this range.

On the right side of Figure 3.13, the relative critical scales were more concentrated within the range of 0.4 to 0.5. The critical scaling magnitude decreased from 2.0 (the smallest size, i.e. highly clustered) to 1.0 (where the distribution becomes a random distribution). The similar decreasing trend was also applied for the final scaling magnitude, which was about 2.75 (highly clustered) to 1.9 (random).

doi:10.6342/NTU201802404

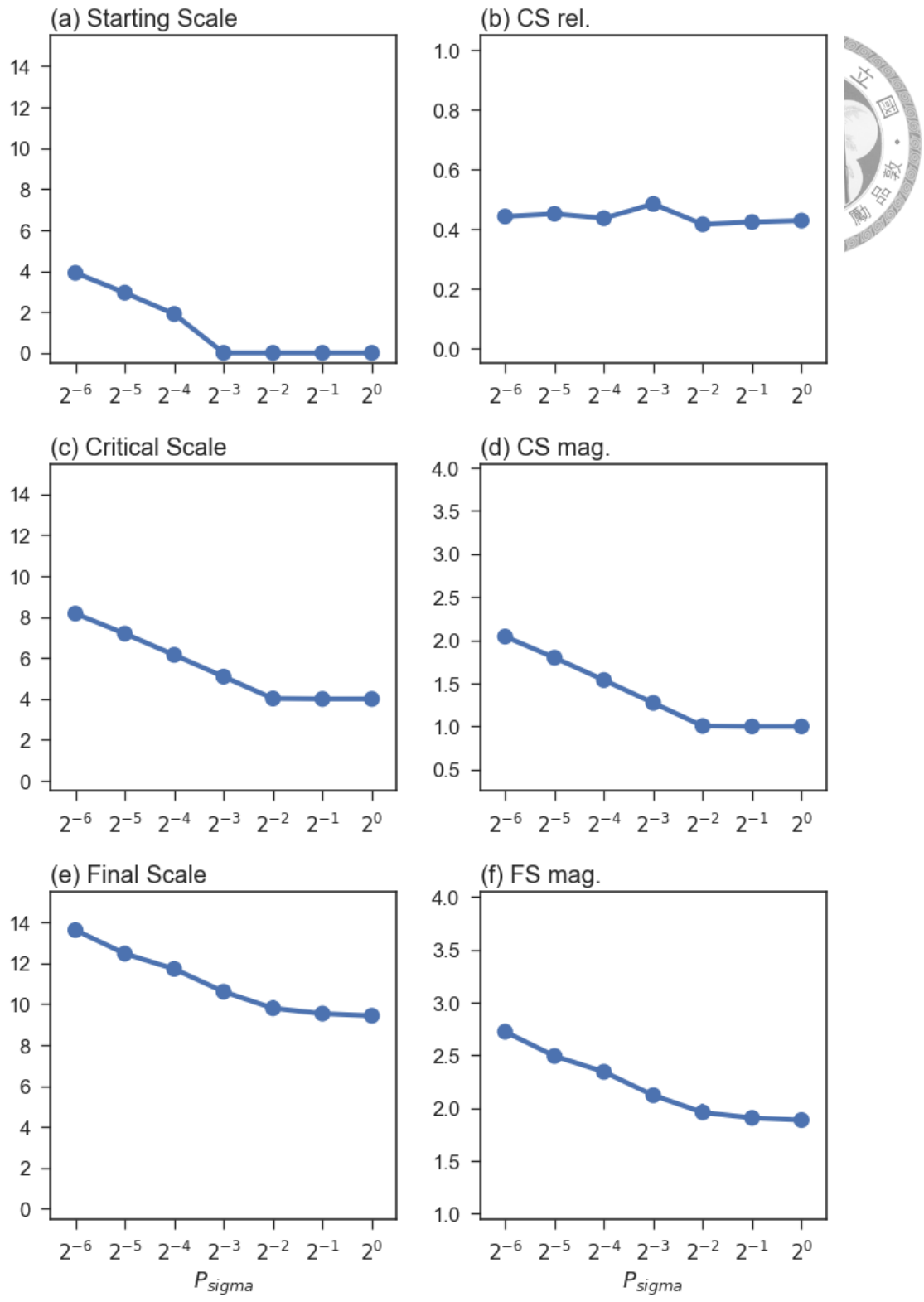


Figure 3.13: The scaling results of part 2 in experiment 2. The changes of the six indexes were shown against the changes of cluster area size, which is starting from smaller ($P_{sigma} = 2^{-6}$) to larger ($P_{sigma} = 2^{0}$).

doi:10.6342/NTU201802404

Table 3.7: The details of the scaling results for part 2 in experiment 2.

P_{sigma}	SS	CS	FS	CS rel.	CS mag.	FS mag.
2^{-6}	3.91(±0.29)	8.17(±0.38)	13.62(±0.96)	0.44(±0.07)	2.04(±0.09)	2.72(±0.19)
2^{-5}	2.94(±0.24)	7.19(±0.4)	12.46(±0.97)	0.45(±0.08)	1.8(±0.1)	2.49(±0.19)
2^{-4}	1.9(±0.44)	6.15(±0.36)	11.72(±1.06)	0.44(±0.07)	1.54(±0.09)	2.34(±0.21)
2^{-3}	0.0(±0.0)	5.09(±0.29)	10.61(±1.01)	0.48(±0.06)	1.27(±0.07)	2.12(±0.2)
2^{-2}	0.0(±0.0)	4.02(±0.14)	9.8(±1.25)	0.42(±0.05)	1.01(±0.04)	1.96(±0.25)
2^{-1}	0.0(±0.0)	4.0(±0.0)	9.54(±0.97)	0.42(±0.04)	1.0(±0.0)	1.91(±0.19)
2^{0}	0.0(±0.0)	4.0(±0.0)	9.43(±1.02)	0.43(±0.04)	1.0(±0.0)	1.89(±0.2)

The part 3 experiment compared the cluster with a different number of points to the scaling analysis result. The starting scales were about two as the degree of separation $P_{sigma} = 2^{-4}$, which showed consistency with the previous part. The critical scale increased from 3 to 7. The final scale, on the other hand, kept increasing from 7 to 14 with a faster-increasing rate than critical scale. As the figure showed that the critical scale and final scale was a linear form compared to the increment of the exponent of two, this suggested that they were the function of log_2 of the number of points. The number of nodes was set to be the increase as the exponent of two, which was starting from 2^4 to 2^{11}, i.e. the exponent was increased by seven times. The final scale was also increased seven level of scales (depth). This might because of the slow increment of the depth for PR-Qtree, which on average case, the increment was about $log_2 N$ (on extremely balance case $log_4 N$).

The three indexes on the right of Figure 3.14 showed a similar finding as previous parts. The relative critical scale concentrated at 0.5 (i.e. the standard deviation is smaller). The critical scaling magnitude and final scaling magnitude were divided by their counterpart of which was a regularly gridded distribution with the same number of points. The critical scaling magnitude decrease from 3.5 to about 1.5, whereas the final scaling magnitude was dropping from about 3.5 to 2.5.

doi:10.6342/NTU201802404

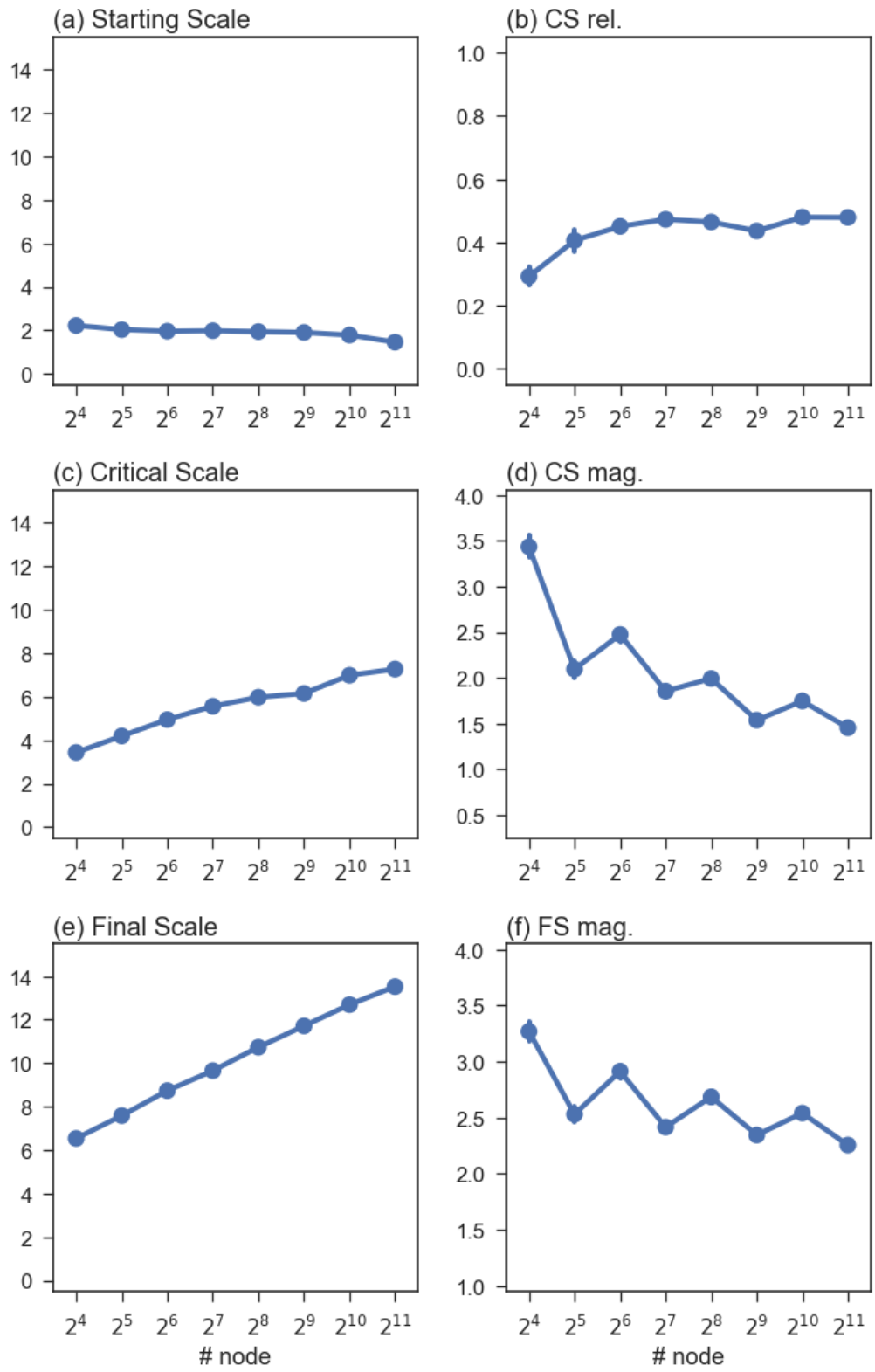


Figure 3.14: The scaling results of part 3 in experiment 2. The changes of the six indexes were shown against the changes of number of points, which is starting from less ($N = 2^4 = 16$) to more ($N = 2^{11} = 2048$).

doi:10.6342/NTU201802404

Table 3.8: The details of the scaling results for part 3 in experiment 2.

no. node	SS	CS	FS	CS rel.	CS mag.	FS mag.
2^4	2.22(±0.42)	3.43(±0.64)	6.54(±0.9)	0.29(±0.14)	3.43(±0.64)	3.27(±0.45)
2^5	2.03(±0.3)	4.19(±0.92)	7.59(±1.03)	0.41(±0.19)	2.1(±0.46)	2.53(±0.34)
2^6	1.96(±0.28)	4.95(±0.75)	8.74(±0.98)	0.45(±0.1)	2.47(±0.37)	2.91(±0.33)
2^7	1.98(±0.2)	5.57(±0.5)	9.66(±0.89)	0.47(±0.09)	1.86(±0.17)	2.41(±0.22)
2^8	1.94(±0.34)	5.98(±0.2)	10.74(±1.0)	0.46(±0.06)	1.99(±0.07)	2.68(±0.25)
2^9	1.9(±0.44)	6.15(±0.36)	11.72(±1.06)	0.44(±0.07)	1.54(±0.09)	2.34(±0.21)
2^{10}	1.78(±0.63)	6.99(±0.1)	12.7(±1.01)	0.48(±0.05)	1.75(±0.03)	2.54(±0.2)
2^{11}	1.45(±0.9)	7.27(±0.45)	13.53(±0.93)	0.48(±0.07)	1.45(±0.09)	2.25(±0.15)

The fourth part of the experiment was to compare the random distribution with varying number of points. The results of starting, critical, and final scales showed similar changing trend finding as to the clustered part (part 3) with a lower value. The critical scale was increased from about 4 to 6, whereas the final scale was increased from about 9 to 13. In other words, the effect of the number of points to the scales was not influenced by the size and shape of the clustering pattern.

Based on the figures on the right of Figure 3.15, the relative critical scale was stable at about 0.4; the critical scaling magnitude was about 1.5 to 1.0; the final scaling magnitude was about 1.9.

Table 3.9: The details of the scaling results for part 4 in experiment 2.

no. node	SS	CS	FS	CS rel.	CS mag.	FS mag.
2^4	0.0(±0.0)	1.8(±0.4)	4.39(±0.9)	0.43(±0.14)	1.8(±0.4)	2.2(±0.45)
2^5	0.0(±0.0)	2.1(±0.3)	5.45(±0.94)	0.4(±0.1)	1.05(±0.15)	1.82(±0.31)
2^6	0.0(±0.0)	2.97(±0.17)	6.53(±1.16)	0.47(±0.08)	1.48(±0.09)	2.18(±0.39)
2^7	0.0(±0.0)	3.02(±0.14)	7.49(±1.14)	0.41(±0.06)	1.01(±0.05)	1.87(±0.28)
2^8	0.0(±0.0)	4.0(±0.0)	8.36(±0.95)	0.48(±0.05)	1.33(±0.0)	2.09(±0.24)
2^9	0.0(±0.0)	4.0(±0.0)	9.43(±1.02)	0.43(±0.04)	1.0(±0.0)	1.89(±0.2)
2^{10}	0.0(±0.0)	5.0(±0.0)	10.36(±1.03)	0.49(±0.04)	1.25(±0.0)	2.07(±0.21)
2^{11}	0.0(±0.0)	5.0(±0.0)	11.48(±1.03)	0.44(±0.04)	1.0(±0.0)	1.91(±0.17)

doi:10.6342/NTU201802404

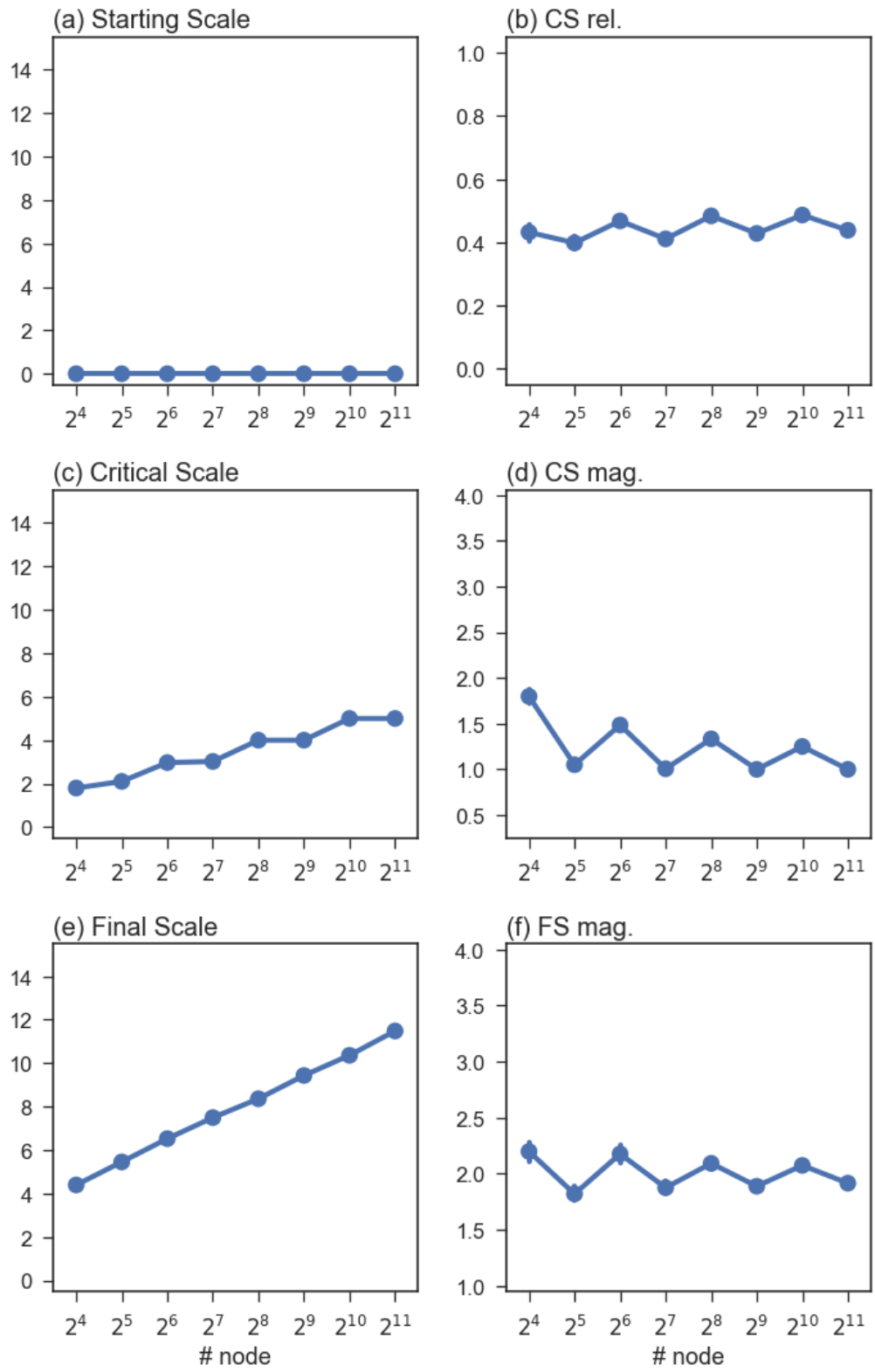
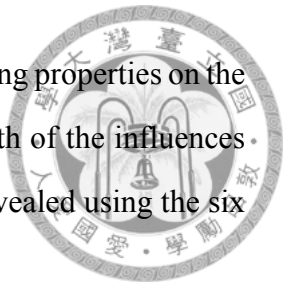


Figure 3.15: The scaling results of part 4 in experiment 2. The changes of the six indexes were shown against the changes of number of points, which is starting from less ($N = 2^4 = 16$) to more ($N = 2^{11} = 2048$).

doi:10.6342/NTU201802404

3.2.5 Summary



In the second experiment, the objective was to test the effect of clustering properties on the scaling analysis results. Through the four parts of studies, the strength of the influences of the locations, covering area size, and the number of points were revealed using the six indexes. The summary of the result is listed as below:

1. The starting scale is sensitive only to the distribution with a very small cover area, that is locally distributed at only one part of the study area.

2. The critical and final scales will change together (maybe with different rates, but will be on the same direction) with the increasing density, which was demonstrated in part 2 in which the number of points was fixed but the area size of the cluster was changed. The linear changing rates indicates that the effect of area size to the critical and final scale is a function of log_2 (Upton and Fingleton, 1979; **?**). On the other hand, the effect of the number of points is also showing a relationship with the critical scale and final scale, which is also a function of log_2.

3. The relative critical scale is always between 0.4 and 0.6 while the distribution is clustered or random (but not regularly distributed as lattice shape, which was described in the first experiment).

4. When the points are distributed randomly, the critical scaling magnitude is equal to 1.0, and the final scaling magnitude is about 1.9. These were shown in part 2 and part 4.

5. When the points are forming a cluster, the critical scaling magnitude will be larger than 1.0, and the final scaling magnitude will be larger than 1.9. These were shown mainly in part 2, and the consistent findings were shown in part 1 and part 3.

doi:10.6342/NTU201802404

3.3 Experiment three: Empirical cases

3.3.1 Aims

In the third experiment, first, this study demonstrates the usage and analysis results using the perspective of point scaling on real-world data; second, this study illustrates the analysis of the spatial macro- and micro-patterns. Three cases were used in this experiments, which (visually speaking) showed different distribution patterns, that were corresponded with the experiment one.

The previous two experiments focused on the theoretical distribution that the clustering properties can be controlled or spatial pattern was known, as they were designed experimentally. But, the real-world point distribution is more complicated due to the fact that they were influenced by a lot of factors related to the nature of the point's underlying context. Therefore, the third experiment focused on the empirical point distribution to run tests of the scaling behaviors using the scaling analysis framework.

3.3.2 Cases dataset

Three cases were used in this experiment, which data were obtained from a POI database (Kingway v18 (2010) 1:5000 digital map), and the Economic Geographic Information System (EGIS) from the Ministry of Economic Affairs, Taiwan (accessed through API from https://egis.moea.gov.tw/). The three cases were post offices (POI database), photocopy shops (EGIS), and beverage shops (EGIS). The study area was set to a joint part of Taipei City and New Taipei City, which contained the densely populated areas within the Taipei Metropolitan Area. The bounding box of the study area was a square shape ($7km \times 7km$), that the horizontal coordinates (east-west direction) were range from $298.500km$ to $305.500km$, and the vertical coordinates (north-south direction) were range from $2766.500km$ to $2773.500km$ (under project EPSG:3826). The point distribution is shown in Figure 3.16.

The point distribution of the empirical cases resembles the three cases in the first experiment (section 3.1). The first case (post office) is sparsely distributed due to the needs of covering the most area. The photocopy shops are concentrated mainly at the south-

doi:10.6342/NTU201802404

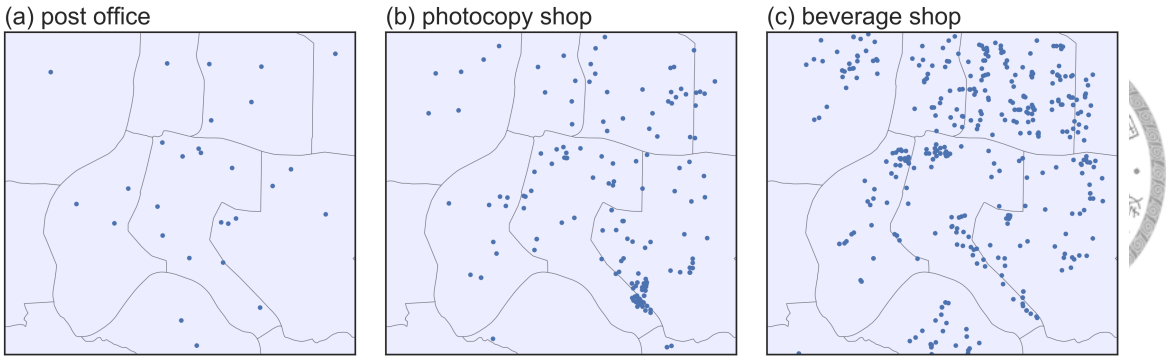


Figure 3.16: The point distributions of the three empirical cases. The points were projected to EPSG:3826.

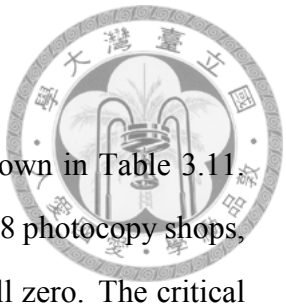
eastern part of the study area, where is the location with a high concentration of schools and universities (i.e. the educational zone in Da'an District); in other words, the high concentration of the demands of photocopy shops is around the area. Having a beverage after meals is a habit for all ages of people in Taiwan, thus beverage shops are common in Taiwan due to high demands, especially at the places with high population. Thus the distribution of beverage shops falls in all zones of the study area while the study area is set to where the population density is high; and which may form a clustering or random distribution. The nearest neighbor analysis of the three cases is shown in Table 3.10, which showed that the case one did not reach the significant level (p-value>0.2), hence it is not a clustering pattern (i.e. random); case two and case three appear to be significant clustering distributions (p-value<0.005). While NNA tests the tendency of clustering, it can only tell if a pattern is significantly clustering (or else random); it does not test the tendency of dispersion or regular.

Table 3.10: The nearest neighbor analysis of the cases in experiment 3.

NNA index	case 1	case 2	case 3
Observed Mean Distance (m)	740.762	181.381	92.633
Expected Mean Distance (m)	661.438	273.304	172.433
Nearest Neighbor Ratio	1.120	0.664	0.537
z-score	1.214	-8.240	17.971
p-value	0.2247	0.0000	0.0000
Result	random	clustered	clustered

doi:10.6342/NTU201802404

3.3.3 Results



Scaling results

The detail information of the scaling results on the three cases is shown in Table 3.11. The number of points was different, that there were 28 post offices, 148 photocopy shops, and 369 beverage shops in the study area. The starting scales were all zero. The critical scales were two, three, and four for the first, second, and third case, respectively. The final scales were five, ten, and eleven, respectively with the cases. The relative critical scale was lower (0.3) for the second case, and was about 0.4 for the first and third cases. The critical scaling magnitudes of the three cases were all one. The final scaling magnitudes were 1.7, 2.5, and 2.2.

Table 3.11: The scaling results of the cases in experiment 3.

case id	case name	no. points	SS	CS	FS	CS rel.	CS mag.	FS mag.
1	post office	28	0	2	5	0.4	1.0	1.7
2	photocopy shop	148	0	3	10	0.3	1.0	2.5
3	beverage shop	369	0	4	11	0.4	1.0	2.2

Figure 3.17 shows the occupied boxes and occupied ratios against the increasing depth (scales) of the three cases. The three cases showed clear bi-fractal structures as discussed in Chapter 2. While the first empirical case resembles the dispersed case in experiment one, it was not a perfect dispersed distribution; some of the locations had closer points, which may be a spatial outliers situation. The second and third cases resembled the corresponding results in the first experiment, except that the tail part (after critical scale) of the second case was relatively longer than its head part (before critical scale).

Aggregation analysis

Figure 3.18 shows the aggregation points of empirical cases on the critical scale using the mean center approach. Some of the key patterns about the macro distribution were emphasized and kept as shown in the aggregated points. The micro patterns which were not visible and not important in terms of the observation of the macro distribution were simplified by point aggregation.

doi:10.6342/NTU201802404

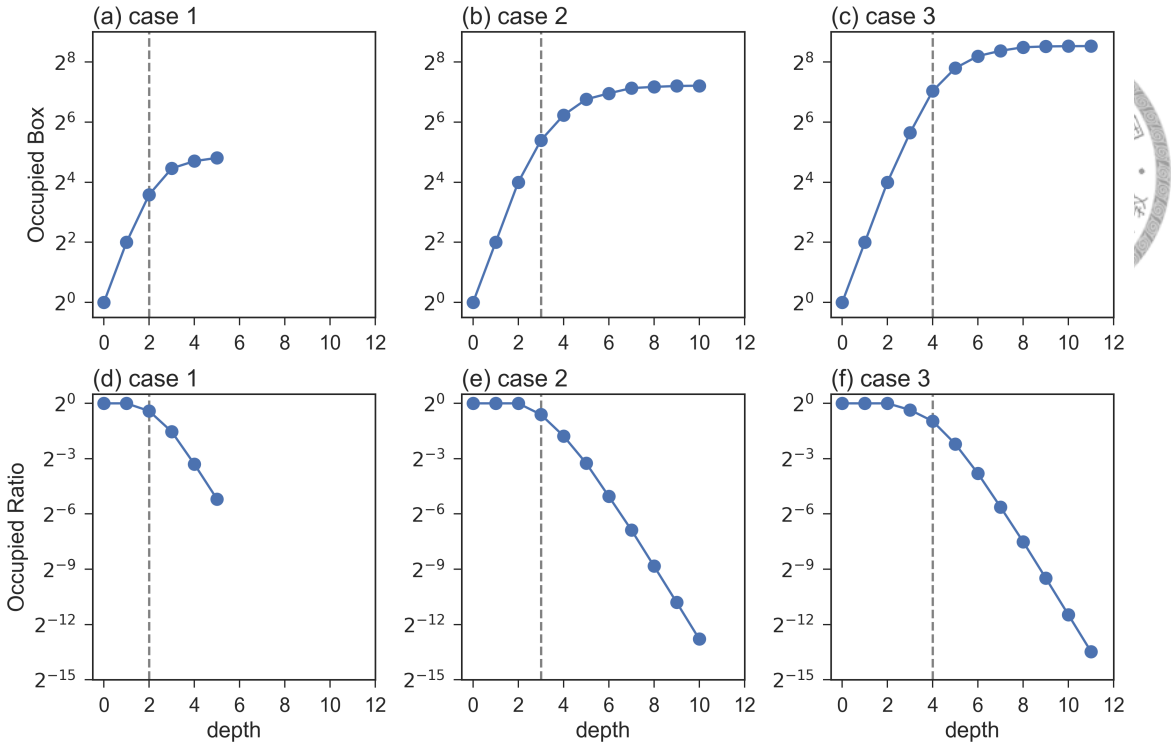


Figure 3.17: The OB-plots and OR-plots of the three empirical cases.

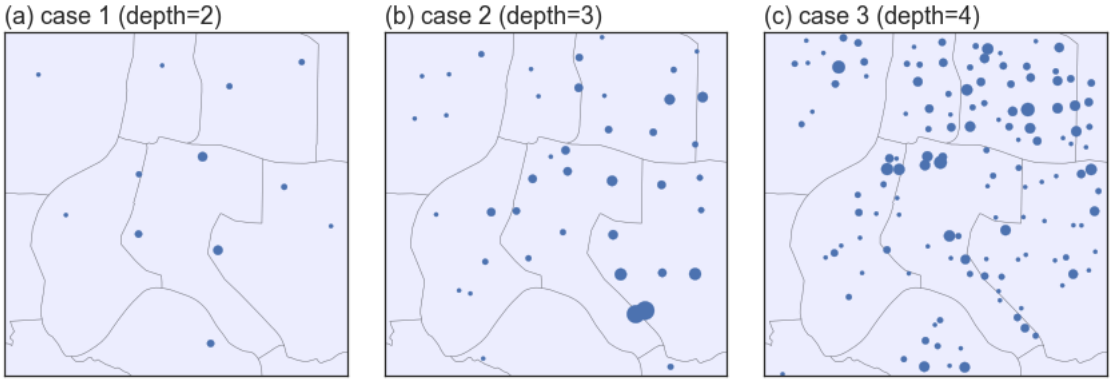


Figure 3.18: The aggregation points of the empirical cases on each critical scale using MC approach.

The RMSE of the five analyses (K-order NNA, K-function, Weighted K-function, KDE, and Weighted KDE), between the original points and aggregation points on each depth (scale), is shown in Figure 3.19. The results showed similar finding as in the first experiment. For the first case, the critical scale was low, and which led to a higher RMSE on critical scale. The performance of critical scales for second and third cases was similar to the corresponding cases in the first experiment, that RMSE of most analyses decreased to a level on the critical scale, especially for the weighted version of analyses.

doi:10.6342/NTU201802404

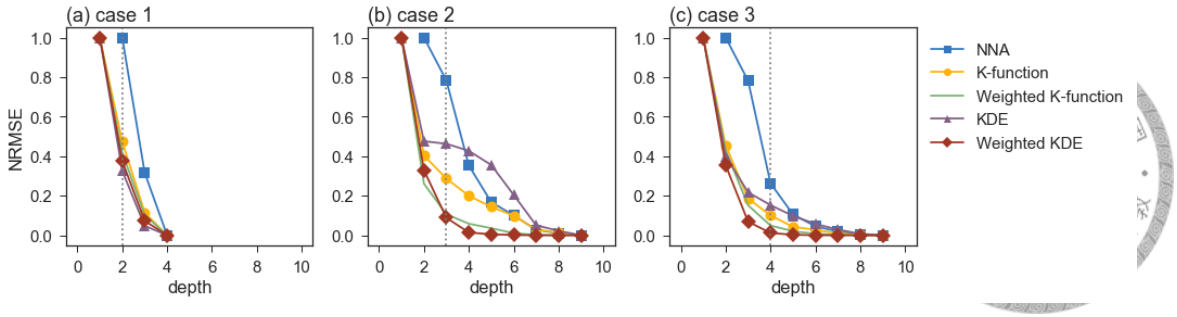


Figure 3.19: The Normalized-Root-Mean-Squared-Error of the five analyses to the aggregation scales.

K-function comparison

To understand the spatial patterns of the aggregation, the K-function and weighted K-function were run for the three cases. The results were shown in Figure 3.20. The K-function of input points were shown in the first column, which showed similar findings as the NNA analysis suggested, that case one was possibly random distribution, case two and three were clustering pattern. The second and third columns showed the analysis result of the spatial pattern using K-function and weighted K-function with the aggregation points on the critical scale.

Based on the K-function analysis, the aggregation points of the first case showed a dispersed distribution. While including the number of point as the weight of each aggregation point, the observed L(d) value was lower than the expected value (blue diagonal line) along with the confidence envelope. Both of the non-weighted and weighted K-function results of the second case suggested that the aggregation points converted to a random distribution from the significant clustering original point distribution. The third case showed similar changes as the second case, that was changed from a significant clustering distribution to the upper bound of the confidence envelope.

doi:10.6342/NTU201802404

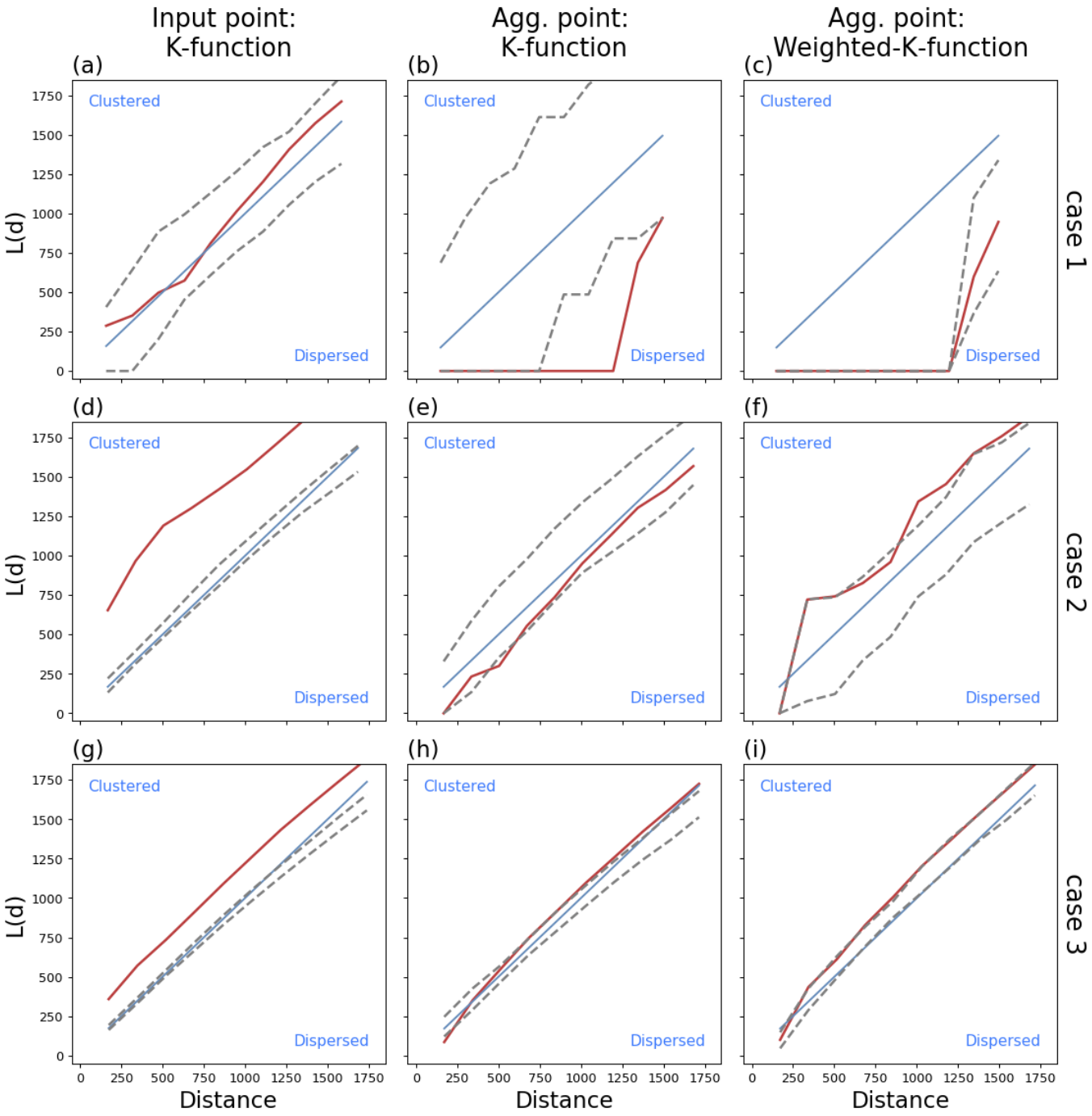


Figure 3.20: The K-function result (L(d)) of the input points and aggregation points: the top row (a-c) for case 1, second row (d-f) for case 2, and bottom row (g-i) for case 3; the left column (a, d, g) is the K-function of input points, middle column (b, e, h) is the (non-weighted) K-function for aggregation points, and the right column (c, f, i) is the Weighted K-function of the aggregation points with the number of merged points as weight.

doi:10.6342/NTU201802404

3.3.4 Summary

In this experiment, the intention is to apply the scaling analysis framework to real-world data and focus on extracting the macro spatial pattern from the original distribution. The results showed that the spatial pattern of empirical data was more complex compared to the theoretical distribution. The summary of findings are described in below:

1. This study used aggregation point distribution to capture the macro spatial pattern. The results suggested that the macro spatial pattern could be different from the original spatial pattern. From the big picture viewpoint, the point distribution may be experiencing one spatial phenomenon, while some of the places may have some location-specific factors that caused the micro distribution to become different. The micro distribution, or spatial outliers, is then be grouped and be considered as the micro spatial pattern. The macro spatial pattern showed the overall distribution of the points that are common in the study area, which is affected by some factors that are happening through all of the areas.

2. While the original point distribution of the post offices (case 1) showed a random pattern, its macro spatial pattern showed a dispersed pattern. This suggested that besides some of the spatial outliers which are close to each other, the rest of the points are forming a dispersed pattern. In other words, these spatial clusters are the reasons that make the distribution becoming random from a dispersed pattern. On the other hand, the location of the spatial outliers is also identified through the aggregated points with higher weight.

3. The second case (photocopy shops) resulted in a clustering pattern while analyzing the original point distribution, but it turned to a random pattern while analyzing the aggregated points (on critical scale level). This finding suggests that if we ignore the locally high-density cluster of photocopy shops that occurs only at a small part of the study area, the rest of the photocopy shops within the study area is randomly distributed. This means that some local factors exist on the cluster location, i.e. the micro pattern, which factors do not necessarily appear at the whole study area or is not a factor that has the same effects on other places in the study area. The result showed that the distribution, i.e. macro pattern, is, therefore, a random pattern if those spatial outliers are ignored. This may because the underlying factors that work the same way within the whole study area may be randomly

doi:10.6342/NTU201802404

distributed, or at least is neither clustered nor dispersed.

4. The third case showed similar results for both original and aggregated distribution, which is a significant clustering pattern (for original point distribution) and slightly significant clustering patterns (for aggregated points on the critical scale). These suggest that the distribution of the original point is forming a similar pattern throughout different scales, i.e. scale-invariant clustering pattern. This result is more like the point distribution in the second experiment, that the appearance at all point may be experiencing the same kind of effect from the underlying spatial processes.

doi:10.6342/NTU201802404

doi:10.6342/NTU201802404

Chapter 4

Discussions

This study designed an analysis framework to identify the critical scale from point distribution based on the PR-Qtree and box-counting method. The framework analyzes the scaling process of the point distribution and reveals the scale that is in between the two fractal dimensions. Then, this study shows a procedure of aggregating points on the critical scale to distinguish the macro pattern and micro pattern of a distribution. The first experiment was conducted to show that the aggregated points on the critical scale were adequate for capturing the overall pattern, i.e. macro pattern, while all of the points were following the same rule for distributing. The second experiment further tested the influences of clustering properties on the scaling results, including the critical scale. The third experiment applied the scaling analysis framework on several real-world distributions, to demonstrates the concept of macro pattern and micro pattern. In the following, integrated issues of the scaling analysis framework were discussed.

4.1 The critical scale of point distribution

The point itself is scale-invariant. Point distribution can be shown in various of scales and possibly form different types of spatial pattern, which was thoroughly discussed in previous MAUP related studies (Openshaw, 1983; Fotheringham and Wong, 1991; Openshaw, 1984; King, 1997; Goodchild, 2011; Taylor et al., 2003; Luoto and Hjort, 2008; Zhu et al., 2001; Zhang et al., 1999; Goodchild et al., 1993). While the spatial pattern of

doi:10.6342/NTU201802404

a fixed set of points is developing through the scales from the most coarse to the finest, this dynamic process follows a trend of changing at the beginning, which was discussed as the fractal dimension of point distribution; and the trend changed after a certain scale – the range starting from which was formally known as the roll-off effect (Agterberg, 2013; Frankhauser, 2015). This means that, on the global level, the scaling issue of the MAUP happens while the scale has changed through the turning point of scales. And this is the concept behind the critical scale of this study.

In other words, the critical scale represents the finest scales that can capture the average pattern of the distribution, i.e. the spatial pattern from a global view and which is not affected by the effect of scaling as discussed in MAUP (or roll-off effect as studied in the related studies of the fractal dimension of points). Therefore, the critical scale and the aggregation points based on the critical scale can be used to visualize or to analyze the spatial point pattern while the aim is to understand the big picture (i.e. the macro pattern) of the point distribution. On the other hand, the critical scale (and final scale) is related to the two components of density (i.e. the number of points and area) as a logarithm function. These finding agreed with previous statistical studies for one dimension data frequency distribution (Silverman, 1986) and multiple categories data (Upton and Fingleton, 1979).

4.2 Data exploration and map visualization

In the age of big data, the exploratory data analysis with a huge number of point event is computationally intensive. While the aggregation point on the critical scale can be used to represent a big picture of the point distribution, it can be used as a surrogate dataset for data exploration. During the aggregation procedure, the point locations that are very close to each other, which distance is not visible on big-picture perspective, is merged and aggregated into a set of representation points to the critical scale level. With a reduced number of points which represents the overall distribution, the aggregated points on the critical scale can reduce the computational resources for complex spatial analyses. Thus, the critical scale and the corresponding aggregation can provide an overall perspective on the data exploration for big data.

doi:10.6342/NTU201802404

Time complexity is a key issue for the calculation framework of a large dataset. The scaling analysis calculation framework is designed based on the Point-Region Quadtree data structure. Theoretically speaking, the height of a PR-Qtree (i.e. the final scale) should be $logN$ for an average case, which may go up to N for the worst case (extremely clustered distribution), or go down to $log_4 N$ for the complete balance case (completely regular/dispersed distribution). The procedures in the calculation, including the construction of PR-Qtree from point location data, the leveling down procedure, and the box-counting procedure, all depend on the number of points (N) and the height of the tree ($logN$). The optimization of the two fractal dimensions depends on the height of the tree (i.e. the range between starting scale and final scale). Therefore, the time complexity of the calculation of the critical scale is about $O(NlogN)$ for average cases. Hence, the computation time of the critical scale should be reasonable, and which results can be helpful for capturing the big picture of the distribution.

On the other hand, point map visualization is also a key problem for cartographer while the number of points is huge. Keeping the distribution pattern while reducing the points into a visible status at the same time is a difficult task. This study provides a framework to aggregate the points into a critical level, which is tested and proved to be suitable for capturing the macro pattern (or the big picture). The aggregation points merged the points that are close and visually overlapped or hard to be differentiated. Thus, mapping them separately can be confusing for visualization. Using the mean center to represent the location of the aggregated points reduces the spatial bias that occurred because of the aggregation. Therefore, the aggregation points on the critical scale should be useful for visualizing the big picture of a huge number of points on a map.

4.3 Macro pattern and micro pattern

Based on the macro pattern, the point distribution is viewed ”as a whole”, i.e. it shows how the points are distributed across the study area. This perspective of viewing the point distribution is described as the **first-order spatial variation** of point distribution (Jiang and Brandt, 2016), which is defined as the **macro pattern** in this study. The concept

doi:10.6342/NTU201802404

of first-order spatial variation intended to describe the underlying factor that is uniform within the area or following a larger trend across the area, and which effect uniformly processes on the events (Cressie, 1993; Getis and Franklin, 1987). While the macro pattern is designed to simplify the spatial anomalies, it is then able to represent the global trends within the study area.

The results from the third experiment showed that the macro pattern of a point distribution may be different from the spatial pattern of the original point data, i.e. the full dataset. These differences exist because some of the points are uncommonly close to each other (as compared to the rest of the point distribution), and these anomaly cases form some spatial outliers within the study area that may dominate or shift the analysis results from the big picture perspective to a dominant view by these anomaly cases. This situation is similar to the situation of one dimension statistical data which frequency distribution is diluted by some outliers.

This study provides a way to represent the macro pattern of a point distribution. The concept of the macro pattern should not be confused with the global patterns or the results of global spatial statistical analyses. The macro pattern is constructed by merging points and weighting the aggregated points to form a more general view about the distribution. In other words, spatially speaking, the location details of points that were too close to each other is ignored; these points will be considered in the weight of the aggregated points. Therefore, the resulting macro pattern is a distribution instead of a spatial statistical analysis (i.e. the global or local analysis). The macro pattern can then be used to calculate further spatial pattern analyses.

Micro pattern, as a part of the macro pattern, is counting the number of points that are merged in the aggregation process and is presented as the weight. The concept of the micro pattern is then related to the second-order spatial variation, which is about the local effects on the occurrences of events. While the weight shows the number of points appear on the surrounding area of an aggregated point, it provides a clear number of how many events happened within the neighboring area. These micro pattern information can then be used for capturing the local variation with some extended analyses.

doi:10.6342/NTU201802404

4.4 The scaling properties

Knowing the critical scale and final scale of a point distribution beforehand can be useful for big data exploration and prediction. On an average case, the critical scale and final scale is usually related to the density of the point distribution, which is a composition of the number of points and area; the results show a logarithm relationship between the density factors and the two scales. This suggested that given a fix study area, the critical scale and final scale should be related to the number of points in the logarithm form.

The results from the extended case studies were shown in Appendix V (with five categories of point events), the final scaling magnitude and the relative critical scale of the cases in the same category are usually falls within a small range. The two extended scaling indexes are designed intended to reduce effects of the number of points to the values of critical scale and final scale. The result of small range variation of the two scaling indexes for a category suggested that although the number of events may be varying across cases (e.g. in different years), their **scaling properties** are similar. Therefore, it is possible to calculate the critical and final scales by using the result from previous case studies (in the same category), with the current estimated number of events. With further analyses and investigations on this aspect, the result can provide a new perspective for data projection or prediction.

4.5 Limitations and future directions

The first issue is about the tested clustering distribution. The clustering model in this study forms point distribution as a single circular pattern which frequency distribution follows a Gaussian function in the horizontal and vertical dimensions. The cases about multiple clusters and non-circular cluster distribution were not tested. These are because of the multiple cluster distribution related to the infinite composition of distribution factors, e.g. the degree of overlapping, the number of clusters, the ratio of points in each cluster, rotation of each cluster, etc. This study used a relatively simple clustering model to test the effects across each factor to investigate the relationship between the clustering effects

doi:10.6342/NTU201802404

to the scaling properties.

The second issue is about the complexity of the real-world case study. The selected cases in the third experiment are also relatively simple with some visually clear spatial pattern. This is because the aim of the experiment is to illustrate the macro and micro pattern from the scaling analysis with a real dataset. Further analyses using more complex and a larger dataset are needed to test the robustness, effectiveness, and time complexity of the scaling analysis framework.

The third issue is about the weighting point distribution. The current scaling analysis framework takes only the location of points into account. The weight of the original point, e.g. the severity of events, the frequency or strength of each event, or the measurement of a sensor station, is ignored in the calculation of the critical scale and also in the aggregation procedure. Therefore, future studies can focus on including the weighting values of the original points in the scaling analysis framework.

The fourth issue is about the dynamic of the events. The dynamical process will cause the PR-Qtree changes across time. Thus, the current framework is not able to be used for real-time analysis; the framework of this study analyzes the events as a time snapshot. To include time-dimension in the analysis framework is a key direction for further development, which may be useful for real-time analysis of the macro pattern and also for the prediction.

doi:10.6342/NTU201802404

Chapter 5

Conclusion

Focusing on the scaling process of point pattern, this study identifies the critical scale which is the finest scale that can capture the big picture of the point spatial distribution. Point pattern will change in the scaling process as mentioned in the modifiable areal unit problem (MAUP) studies (Openshaw, 1983; Fotheringham and Wong, 1991). In the process of scaling from coarse to fine, the spatial point pattern become clearer and more spatial variations details will appear as the resolution become higher. But, this study found that the growth of the spatial variation details will reach to the critical scale level, and the further increment of resolution will not change the spatial pattern at the most area, i.e. the increment of spatial variation details will only happen at some of the locations that have extremely close points. In this study, these small ranges of locations are recognized as the spatial outliers, i.e. the spatial distance between the points in these regions are lower than the ordinary distances between the pair of points at the rest of the whole study area. On the other hand, the distribution pattern on the critical scale is then defined as the macro pattern, i.e. representing the big picture of the point distribution.

In order to differentiate the macro pattern from the whole set of point distribution, this study proposed an analytical framework for identifying the critical scale, which is designed based on the concept of the bi-fractal nature of spatial point distribution (Agterberg, 2013), and the point-region quadtree (PR-Qtree) data structure. In addition, based on the critical scale, the macro pattern is able to be produced and which can represent the big picture of the distribution. The conceptual foundation of the critical scale is based on the conversion

doi:10.6342/NTU201802404

from the scaling range where the increment of details is significant for the whole area, to the scaling range where the increment will only happen at several small parts of areas; this conversion is described as the bi-fractal pattern of point distribution (Frankhauser, 2015). The implementation of the concept is based on the PR-Qtree data structure, which accelerates the computation processes, and provides a foundation for the calculation of scaling process by converting the concept of scale as the depth of the tree. By integrating the bi-fractal concept and PR-Qtree, this study intended to achieve a breakthrough on the scaling aspect of MAUP by reconstructing the macro pattern based on the critical scale of point distribution.

The critical scale is a degree of resolution that is suitable to view the big picture of the point distribution and to derive the spatial patterns. The critical scale divided the point distribution into two scaling range: (1) the macro pattern that view the point distribution as a whole, which is captured by the scaling phenomenon before the critical scale (head part); and (2) the local patterns that occur at some specific locations, which is captured by the scaling phenomenon after the critical scale (tail part). These two scaling ranges were described as the first-order and second-order spatial variation effects (Jiang and Brandt, 2016). The first-order effect focuses on the spatial heterogeneity issue, which is observing the world from the global level, and is describing the spatial patterns by the organization of distribution (e.g. most of the geographical variation is in general spatially wild rather than mild). The second-order effect focuses on the spatial dependence, that is focusing on the local level, and is describing the spatial patterns by the distance from one its neighbors (e.g. that near things are more related than distant things). This study proposed a novel framework for differentiating the two scaling ranges and conducted a series of tests using both theoretical and empirical point distributions.

The main findings from the experiment can be summarized as follows: (a) the aggregation point pattern on critical scales can capture the spatial pattern of the original spatial pattern; (b) the density of a cluster will affect the resulting critical scale and final scale, which relationships can be described using a logarithm function; (c) the macro pattern may not always consistent with the original distribution, that the original spatial pattern

doi:10.6342/NTU201802404

may be dominated by a part of the points which may be a spatial outliers.

Aside from the conventional understanding of point pattern as discussed in the previous global or local spatial statistics studies, this study provides not only a new tool but also a novel perspective of viewing the point distribution. This analysis framework, including the critical scale identification and macro pattern aggregation, can be useful for spatial point data exploration and map visualization.

doi:10.6342/NTU201802404

doi:10.6342/NTU201802404

References

Agterberg, F. P. (2013). Fractals and spatial statistics of point patterns. *Journal of Earth Science*, 24(1):1–11.

Anderson, T. K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, 41(3):359–364.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical analysis*, 27(2):93–115.

Arribas-Bel, D., de Graaff, T., and Rey, S. J. (2017). Looking at John Snow's Cholera map from the twenty first century: A practical primer on reproducibility and open science. In *Regional Research Frontiers-Vol. 2*, pages 283–306. Springer.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive spatial data analysis*, volume 413. Longman Scientific & Technical Essex.

Batty, M. (2007). *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*. The MIT press.

Batty, M. (2008). The size, scale, and shape of cities. *science*, 319(5864):769–771.

Batty, M., Longley, P., and Fotheringham, S. (1989). Urban growth and form: scaling, fractal geometry, and diffusion-limited aggregation. *Environment and Planning A*, 21(11):1447–1472.

Batty, M. and Longley, P. A. (1994). *Fractal cities: a geometry of form and function*. Academic press.

doi:10.6342/NTU201802404

Blenkinsop, T. G. and Sanderson, D. J. (1999). Are gold deposits in the crust fractals? A study of gold mines in the Zimbabwe craton. *Geological Society, London, Special Publications*, 155(1):141–151.

Boots, B. N. and Getis, A. (1988). *Point Pattern Analysis*. Sage Publications, Inc.

Carlson, C. A. (1991). Spatial distribution of ore deposits. *Geology*, 19(2):111–114.

Chainey, S., Reid, S., and Stuart, N. (2002). *When is a hotspot a hotspot? A procedure for creating statistically robust hotspot maps of crime*. Taylor & Francis, London, England.

Chen, Y. and Wang, J. (2013). Multifractal characterization of urban form and growth: the case of Beijing. *Environment and Planning B: Planning and Design*, 40(5):884–904.

Clark, P. J. and Evans, F. C. (1954). Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 35(4):445–453.

Cressie, N. (1993). *Chapter 8: Spatial point patterns*. John Wiley & Sons.

Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. CRC Press.

Feng, J. and Chen, Y. (2010). Spatiotemporal evolution of urban form and land-use structure in Hangzhou, China: evidence from fractals. *Environment and planning B: Planning and design*, 37(5):838–856.

Ford, A. and Blenkinsop, T. G. (2008). Combining fractal analysis of mineral deposit clustering with weights of evidence to evaluate patterns of mineralization: application to copper deposits of the Mount Isa Inlier, NW Queensland, Australia. *Ore Geology Reviews*, 33(3):435–450.

Fotheringham, A. S. and Wong, D. W. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A*, 23(7):1025–1044.

Fotheringham, A. S. and Zhan, F. B. (1996). A comparison of three exploratory methods for cluster detection in spatial point patterns. *Geographical analysis*, 28(3):200–218.

doi:10.6342/NTU201802404

Frankhauser, P. (2004). Comparing the morphology of urban patterns in Europe–a fractal approach. *European Cities–Insights on outskirts, Report COST Action*, 10:79–105.

Frankhauser, P. (2015). From fractal urban pattern analysis to fractal urban planning concepts. In *Computational Approaches for Urban Environments*, pages 13–48. Springer.

Gatrell, A. C., Bailey, T. C., Diggle, P. J., and Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British geographers*, pages 256–274.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61:115–125.

Getis, A. (1984). Interaction modeling using second-order analysis. *Environment and Planning A*, 16(2):173–183.

Getis, A. and Franklin, J. (1987). Second-Order Neighborhood Analysis of Mapped Point Patterns. *Ecology*, 68(3):473–477.

Goodchild, M. F. (2011). Scale in GIS: An overview. *Geomorphology*, 130(1):5–9.

Goodchild, M. F., Anselin, L., and Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and planning A*, 25(3):383–397.

Goodchild, M. F. and Mark, D. M. (1987). The fractal nature of geographic phenomena. *Annals of the Association of American Geographers*, 77(2):265–278.

Gumiel, P., Sanderson, D., Arias, M., Roberts, S., and Martín-Izard, A. (2010). Analysis of the fractal clustering of ore deposits in the Spanish Iberian Pyrite Belt. *Ore Geology Reviews*, 38(4):307–318.

Hayakawa, M., Ito, T., and Smirnova, N. (1999). Fractal analysis of ULF geomagnetic data associated with the Guam earthquake on August 8, 1993. *Geophysical Research Letters*, 26(18):2797–2800.

Jiang, B. and Brandt, S. A. (2016). A fractal perspective on scale in geography. *ISPRS*

doi:10.6342/NTU201802404

International Journal of Geo-Information, 5(6):95.

Jiang, S. and Liu, D. (2012). Box-counting dimension of fractal urban form: stability issues and measurement design. *International Journal of Artificial Life Research (IJALR)*, 3(3):41–63.

Kagan, Y. Y. and Jackson, D. D. (1991). Long-term earthquake clustering. *Geophysical Journal International*, 104(1):117–134.

Keersmaecker, M.-L., Frankhauser, P., and Thomas, I. (2003). Using fractal dimensions for characterizing intra-urban diversity: The example of Brussels. *Geographical analysis*, 35(4):310–328.

King, G. (1997). *A solution to the ecological inference problem: Reconstructing Individual Behavior from Aggregate Data.* Princeton University Press, Princeton, NJ.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in medicine*, 14(8):799–810.

Lawson, A. B. and Denison, D. G. (2002). *Spatial cluster modelling*. CRC press.

Lee, J., Lay, J.-G., Chin, W. C. B., Chi, Y.-L., and Hsueh, Y.-H. (2014). An experiment to model spatial diffusion process with nearest neighbor analysis and regression estimation. *International Journal of Applied Geospatial Research (IJAGR)*, 5(1):1–15.

Luoto, M. and Hjort, J. (2008). Downscaling of coarse-grained geomorphological data. *Earth Surface Processes and Landforms*, 33(1):75–89.

Mandelbrot, B. (1967). How long is the coast of Britain? Statistical self-similarity and fractional dimension. *science*, 156(3775):636–638.

McCord, E. S. and Ratcliffe, J. H. (2009). Intensity value analysis and the criminogenic effects of land use features on local crime patterns. *Crime Patterns and Analysis*, 2(1):17–30.

doi:10.6342/NTU201802404

Nakaya, T. and Yano, K. (2010). Visualising Crime Clusters in a Space-time Cube: An Exploratory Data-analysis Approach Using Space-time Kernel Density Estimation and Scan Statistics. *Transactions in GIS*, 14(3):223–239.

Openshaw, S. (1983). *The Modifiable Areal Unit Problem.* GeoBooks, Norwick, UK.

Openshaw, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and planning A*, 16(1):17–31.

Ord, J. K. and Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis*, 27(4):286–306.

Orenstein, J. A. (1982). Multidimensional tries used for associative searching. *Information Processing Letters*, 14(4):150–157.

Pickering, G., Bull, J., and Sanderson, D. (1995). Sampling power-law distributions. *Tectonophysics*, 248(1-2):1–20.

Raines, G. L. (2008). Are fractal dimensions of the spatial distribution of mineral deposits meaningful? *Natural Resources Research*, 17(2):87.

Samet, H. (1984). The quadtree and related hierarchical data structures. *ACM Computing Surveys (CSUR)*, 16(2):187–260.

Seaman, D. E. and Powell, R. A. (1996). An evaluation of the accuracy of kernel density estimators for home range analysis. *Ecology*, 77(7):2075–2085.

Sémécurbe, F., Tannier, C., and Roux, S. G. (2016). Spatial distribution of human population in France: Exploring the modifiable areal unit problem using multifractal analysis. *Geographical Analysis*, 48(3):292–313.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Chapman and Hall/CRC.

Sutton, P. C. (2003). A scale-adjusted measure of "urban sprawl" using nighttime satellite imagery. *Remote sensing of environment*, 86(3):353–369.

doi:10.6342/NTU201802404

Tannier, C. and Pumain, D. (2005). Fractals in urban geography: a theoretical outline and an empirical example. *Cybergeo: European Journal of Geography*.

Taylor, C., Gorard, S., and Fitz, J. (2003). The modifiable areal unit problem: segregation between schools and levels of analysis. *International Journal of Social Research Methodology*, 6(1):41–60.

Terzi, F. and Kaya, H. S. (2011). Dynamic spatial analysis of urban sprawl through fractal geometry: the case of Istanbul. *Environment and Planning B: Planning and Design*, 38(1):175–190.

Thomas, I., Frankhauser, P., and Biernacki, C. (2008). The morphology of built-up landscapes in Wallonia (Belgium): A classification using fractal indices. *Landscape and urban planning*, 84(2):99–115.

Upton, G. and Fingleton, B. (1979). Log-linear models in geography. *Transactions of the Institute of British Geographers*, pages 103–115.

Upton, G., Fingleton, B., et al. (1985). *Chapter 1: The identification of patterns*. John Wiley & Sons Ltd.

Varnes, D. J. and Bufe, C. G. (1996). The cyclic and fractal seismic series preceding an mb 4.8 earthquake on 1980 February 14 near the Virgin Islands. *Geophysical Journal International*, 124(1):149–158.

Venables, W. N. and Ripley, B. D. (2002). *Tree-Based Methods*, pages 251–269. Springer New York, New York, NY.

Walsh, J., Watterson, J., and Yielding, G. (1991). The importance of small-scale faulting in regional extension. *Nature*, 351(6325):391.

White, R. and Engelen, G. (1993). Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land-use patterns. *Environment and planning A*, 25(8):1175–1199.

doi:10.6342/NTU201802404

White, R., Engelen, G., and Uljee, I. (2015). *Modeling cities and regions as complex systems: From theory to planning applications*. MIT Press.

Wong, W. and Lee, J. (2005). *Statistical analysis of geographic information with ArcView GIS and ArcGIS*. Wiley.

Yu, X. J. and Ng, C. N. (2007). Spatial and temporal dynamics of urban sprawl along two urban–rural transects: A case study of Guangzhou, China. *Landscape and Urban Planning*, 79(1):96–109.

Zhang, X., Drake, N. A., Wainwright, J., and Mulligan, M. (1999). Comparison of slope estimates from low resolution DEMs: Scaling issues and a fractal method for their solution. *Earth Surface Processes and Landforms*, 24(9):763–779.

Zhu, A.-X., Hudson, B., Burt, J., Lubich, K., and Simonson, D. (2001). Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Science Society of America Journal*, 65(5):1463–1472.

doi:10.6342/NTU201802404

doi:10.6342/NTU201802404

Appendix I: Model for generating clustering distribution

A Mono-centric Clustering Model

To generate clustered points, a simple, mono-centric clustering model is used. In this model, a 2-dimensions Gaussian distribution is used as the probability function for generating the points (Lawson and Denison (2002)). In other words, the x- and y-coordinates of a random point were generated separately, and which locations (values) are both following the Gaussian distribution. A mean(μ) and a standard deviation ($sigma$) are needed to formulate a Gaussian distribution probability function. In the generation of random cluster points, the mean is the location of the cluster center, while the standard deviation determines the shape of the Gaussian distribution. The sigma value is converted using Equation 1 to the proportion of sigma (P_{sigma}) in comparison to the whole length (L) of the side of the study area. While the standard deviation of the Gaussian can be used generate the random value with different shape of the Gaussian distribution, it can be used to control the separation of points from the mean value (i.e. center of the distribution). Thus, this study uses the P_{sigma} to control the degree of separation of the clustering phenomenon. Figure 1 shows the generated coordinate using different P_{sigma}. The P_{sigma} on the left (a) is the lowest, resulting in a more concentrated coordinates at the center (i.e. μ); the P_{sigma} on the middle (b) is slightly higher than the P_{sigma} on the left, resulting in a slightly flatten of the frequency distribution; the P_{sigma} on the right (c) is the highest

doi:10.6342/NTU201802404

among the three, the histogram showed a uniform pattern.

$$P_{sigma} = \frac{\sigma}{L} \tag{1.1}$$

$$\sigma = L \times P_{sigma} \tag{1.2}$$

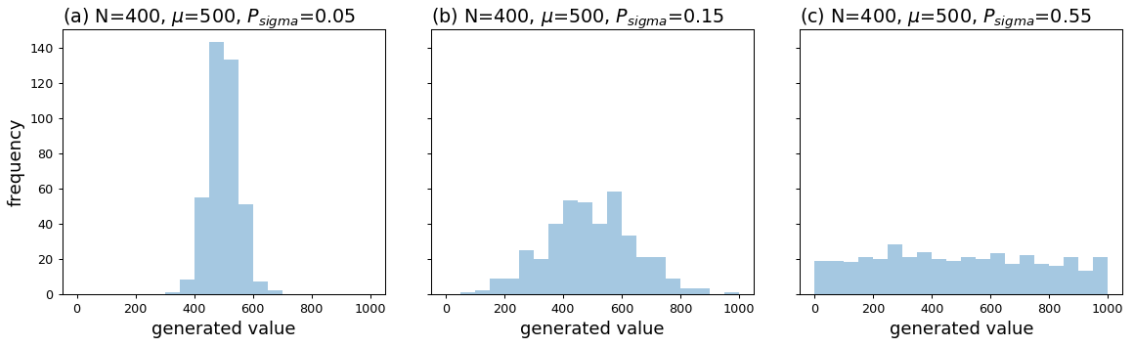


Figure 1: Three distributions of 400 random points generated using Gaussian distribution with $\mu = 500$ and different sigma values: (a) $P_{sigma} = 0.05$, or $sigma = 50$; (b) $P_{sigma} = 0.15$, or $sigma = 150$; (c) $P_{sigma} = 0.55$, or $sigma = 550$.

Using P_{sigma}, the intensity of concentration of the clusters can be controlled. Therefore, the mono-centric clustering point model used the 2-dimension Gaussian distribution, that is, generating x- and y- coordinates independently, to generate enough number of points, which is also another parameter that controls the density. The model is designed as shown in Algorithm 4, which also included a checking procedure to confirm the all of the points generated are still within the x- and y-range.

A series of examples of generated point distribution using different separation parameters and center coordinates were shown in Figure 2

doi:10.6342/NTU201802404

Algorithm 4 Mono-centric clustering point model

procedure Mono-centric($N, C(x_c, y_c), P_{sigma}, Box(x_0, y_0, x_1, y_1)$)
  $points \leftarrow List[N]$
  $sigma_x \leftarrow P_{sigma} \times (x1 - x0)$
  $sigma_y \leftarrow P_{sigma} \times (y1 - y0)$
  $i \leftarrow 0$
  while $i < N$ **do**
    $x \leftarrow RandomGauss(\mu = x_c, \sigma = sigma_x)$ ▷ μ & σ as mean and std
    $y \leftarrow RandomGauss(\mu = y_c, \sigma = sigma_y)$
    if ($x_0 \leq x \leq x_1$) & ($y_0 \leq y \leq y_1$) **then**
      $points[i] \leftarrow P(x, y)$
      $i+ = 1$
    end if
  end while ▷ enough points were generated
  return $points$
end procedure

doi:10.6342/NTU201802404

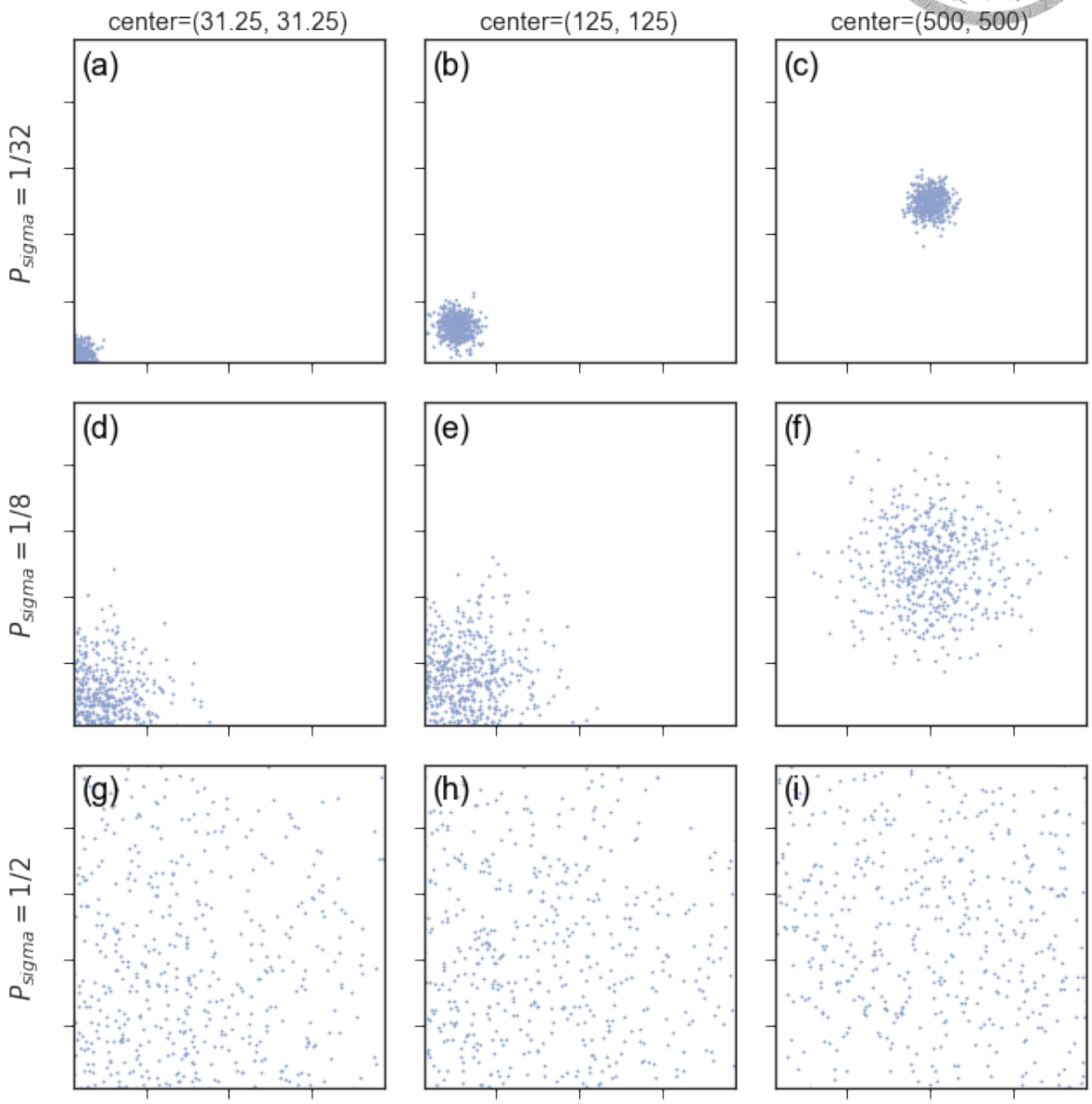


Figure 2: The example of points generated using mono-centric clustering model with different combination of P_{sigma} and center of cluster.

doi:10.6342/NTU201802404

Appendix II: Analysis process of experiment one

In this appendix, the outputs of five analyses, including the three global pattern analyses (NNA, K-function, and Weighted K-function) and the two local pattern analyses (KDE and Weighted KDE), were presented. These outputs were used to calculate the root-mean-square-error (RMSE), which will be converted to a normalized RMSE (Equation 2) for cross-analyses comparison purpose.

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \tag{2}$$

doi:10.6342/NTU201802404

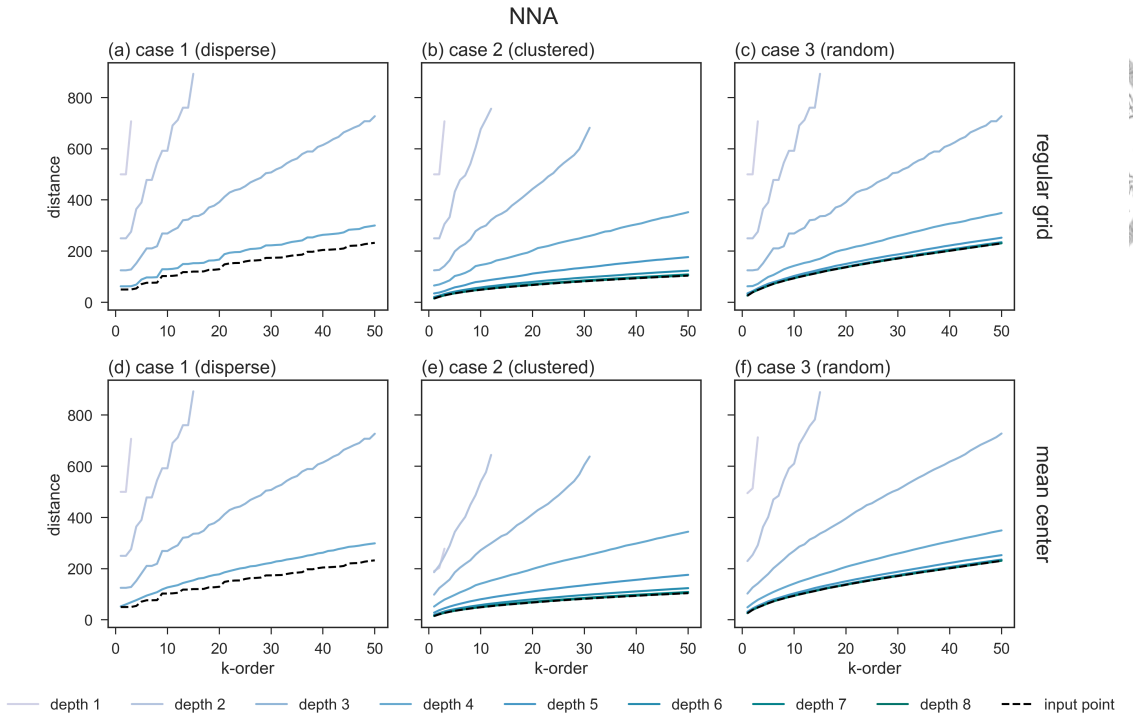


Figure 3: The NNA result of the three cases. The sub-figures in the above row (a, b, c) are calculated using grid center approach, whereas the sub-figures on the below (d, e, f) are calculated using mean center approach.

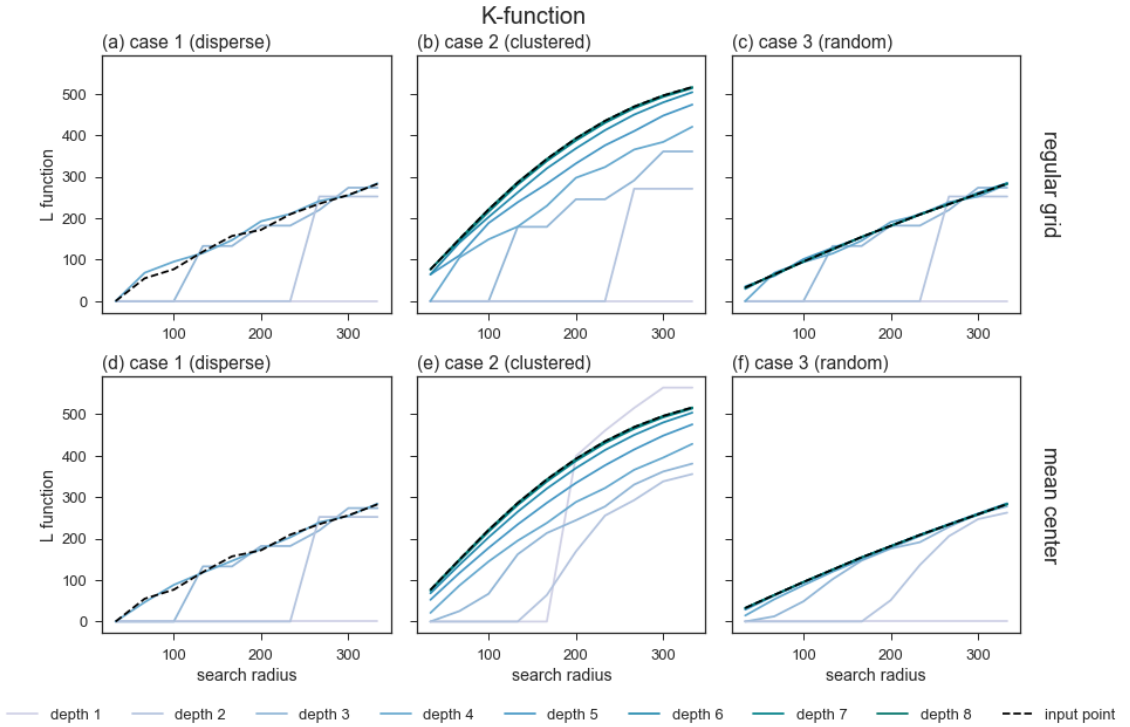


Figure 4: The K-function result of the three cases. The sub-figures in the above row (a, b, c) are calculated using grid center approach, whereas the sub-figures on the below (d, e, f) are calculated using mean center.

doi:10.6342/NTU201802404

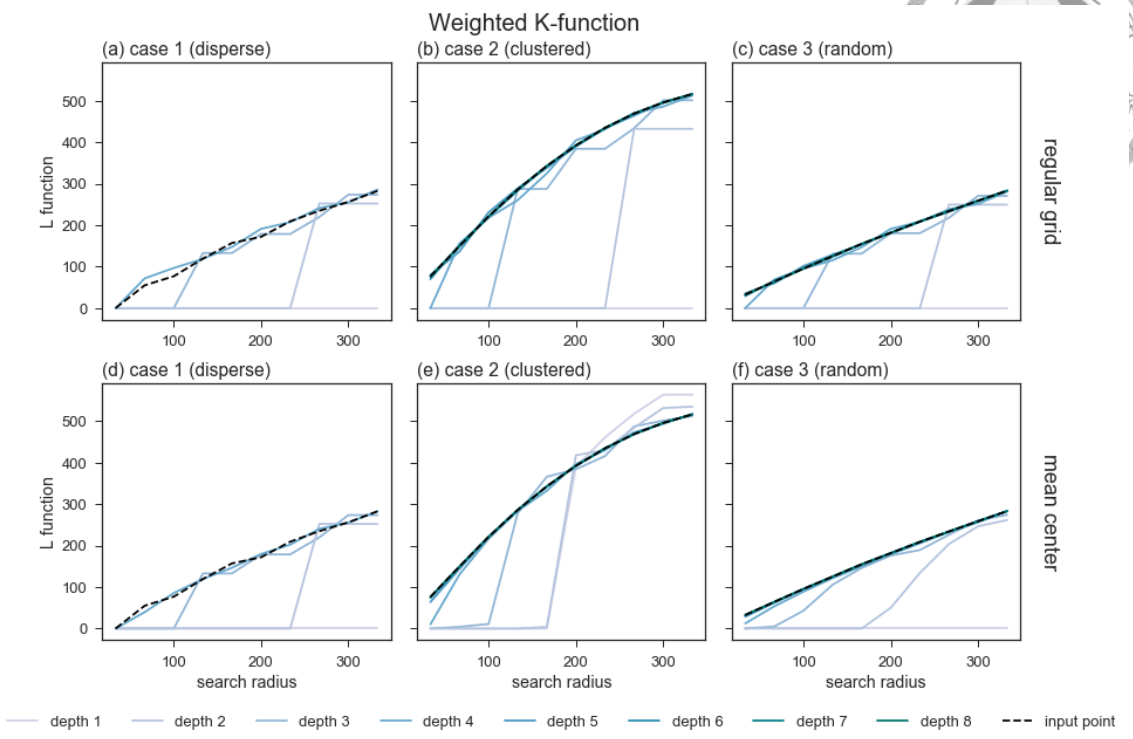


Figure 5: The Weighted K-function result of the three cases. The sub-figures in the above row (a, b, c) are calculated using grid center approach, whereas the sub-figures on the below (d, e, f) are calculated using mean center approach.

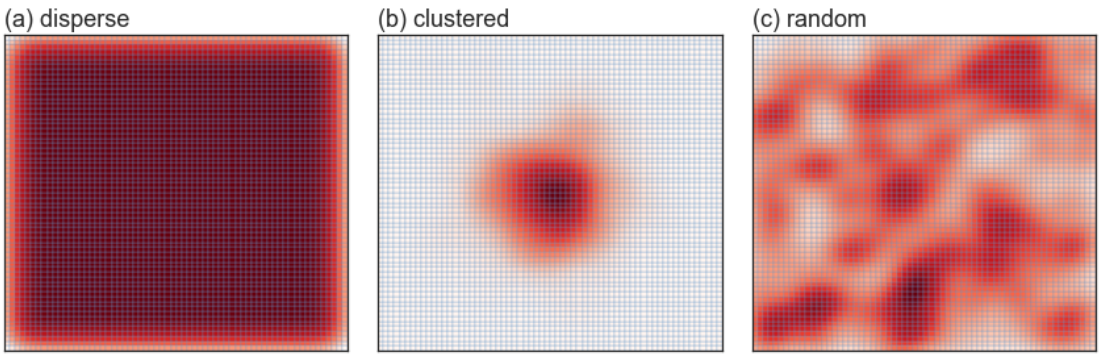


Figure 6: The KDE of the original input points of the three cases. While each point indicates an event, it is weighted as one. Therefore the weighted and unweighted KDE will have the same output results. This is the comparison bases for calculating the RMSE.

doi:10.6342/NTU201802404

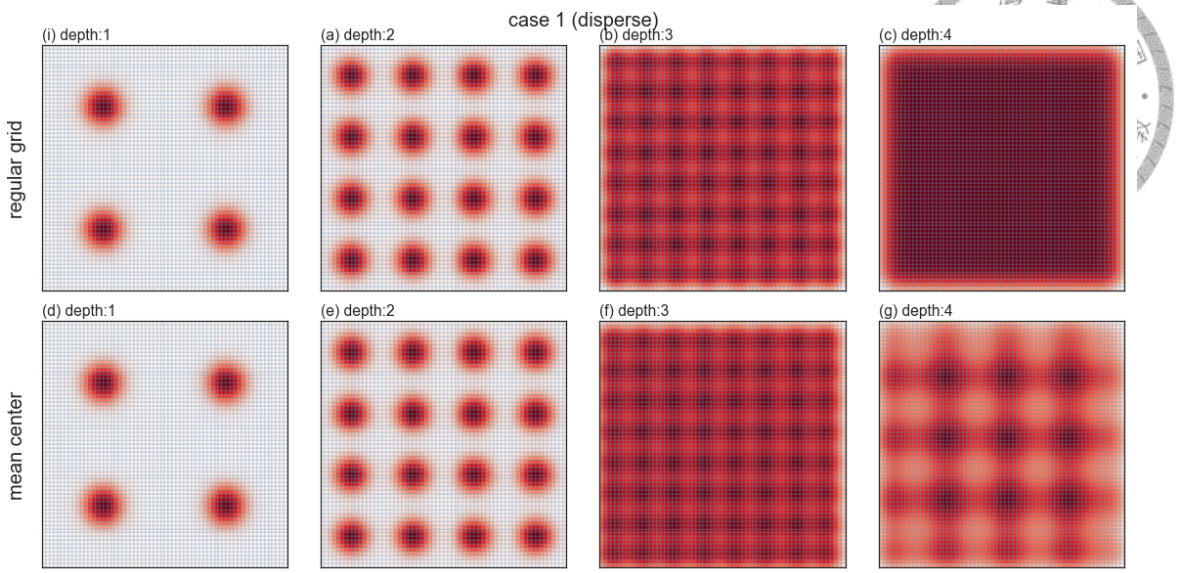


Figure 7: The KDE of case 1 without considering the counting weight. The sub-figures in the above row (a, b, c, d) are calculated using grid center approach, whereas the sub-figures on the below (e, f, g, h) are calculated using mean center approach. Each sub-figure in a row indicates the aggregation using a depth, starting from one to four.

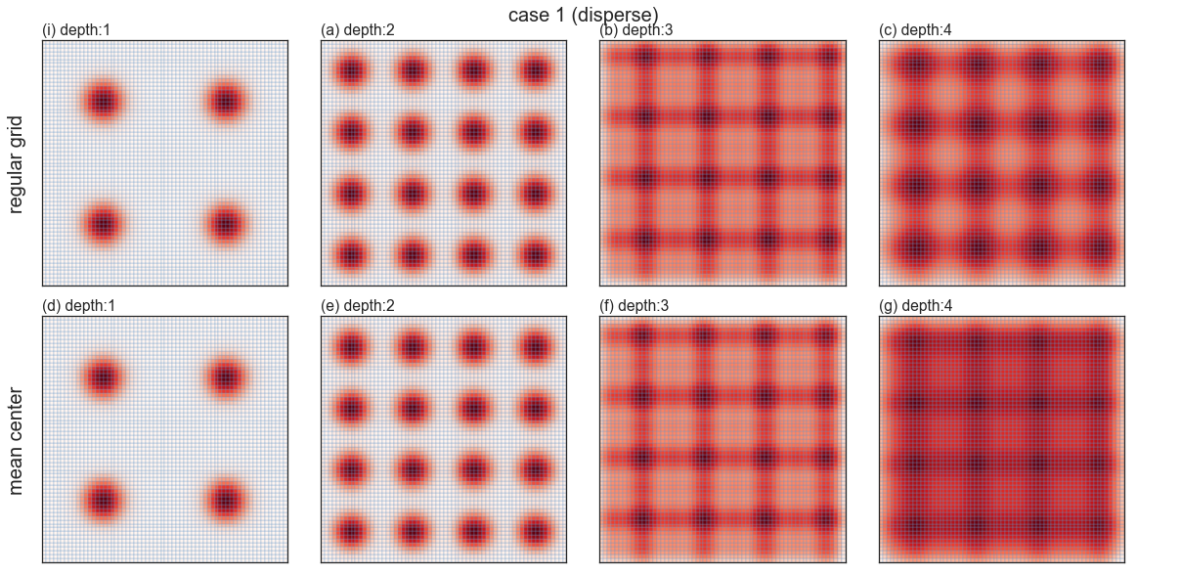


Figure 8: The Weighted KDE of case 1. The sub-figures in the above row (a, b, c, d) are calculated using grid center approach, whereas the sub-figures on the below (e, f, g, h) are calculated using mean center approach. Each sub-figure in a row indicates the aggregation using a depth, starting from one to four.

doi:10.6342/NTU201802404

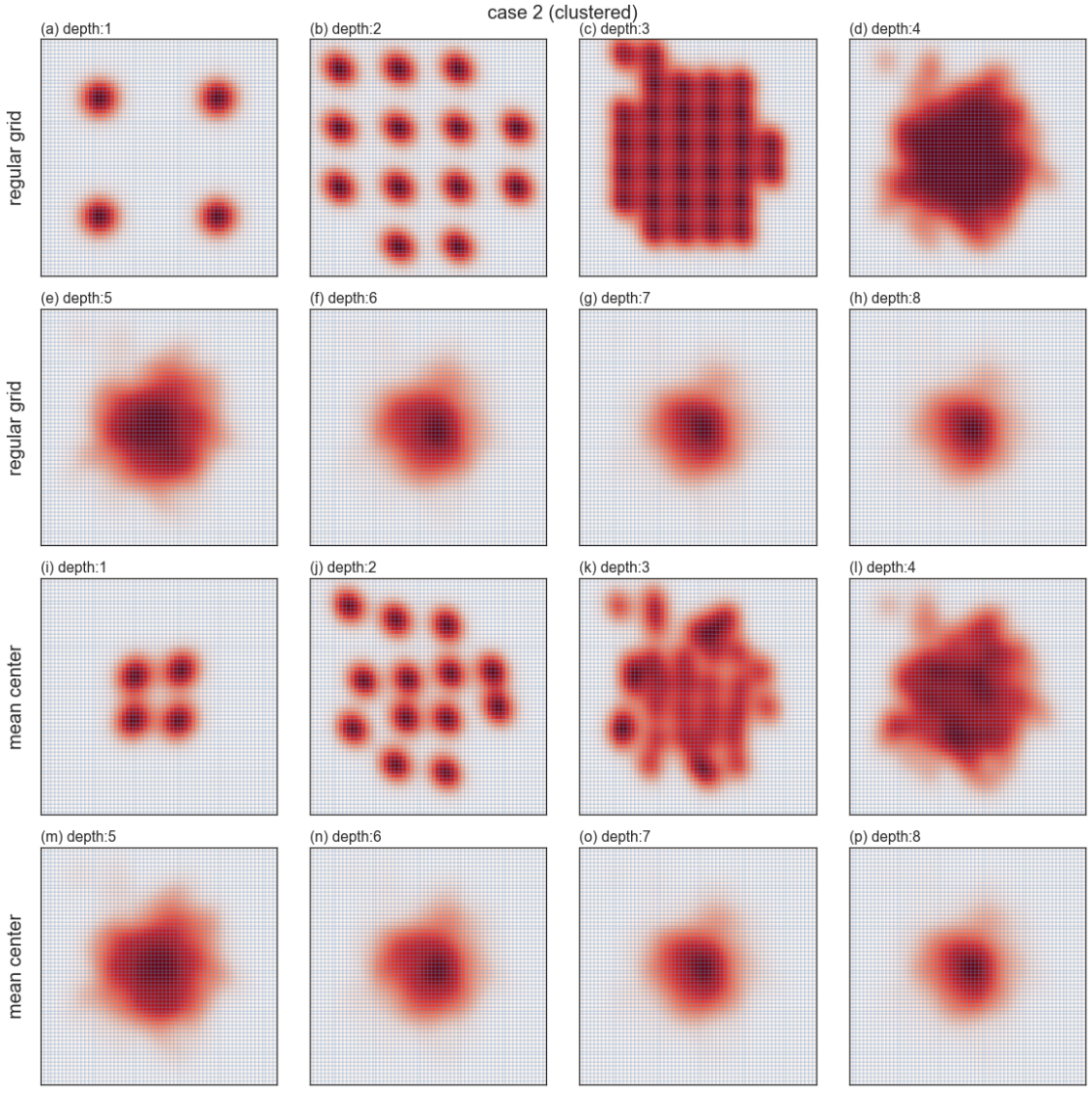


Figure 9: The KDE of case 2 without considering the counting weight. The sub-figures in the first two rows (a-h) are calculated using grid center approach, whereas the sub-figures on the last two rows (i-p) are calculated using mean center approach. Each sub-figure within the rows indicates the aggregation using a depth, starting from one to eight.

doi:10.6342/NTU201802404

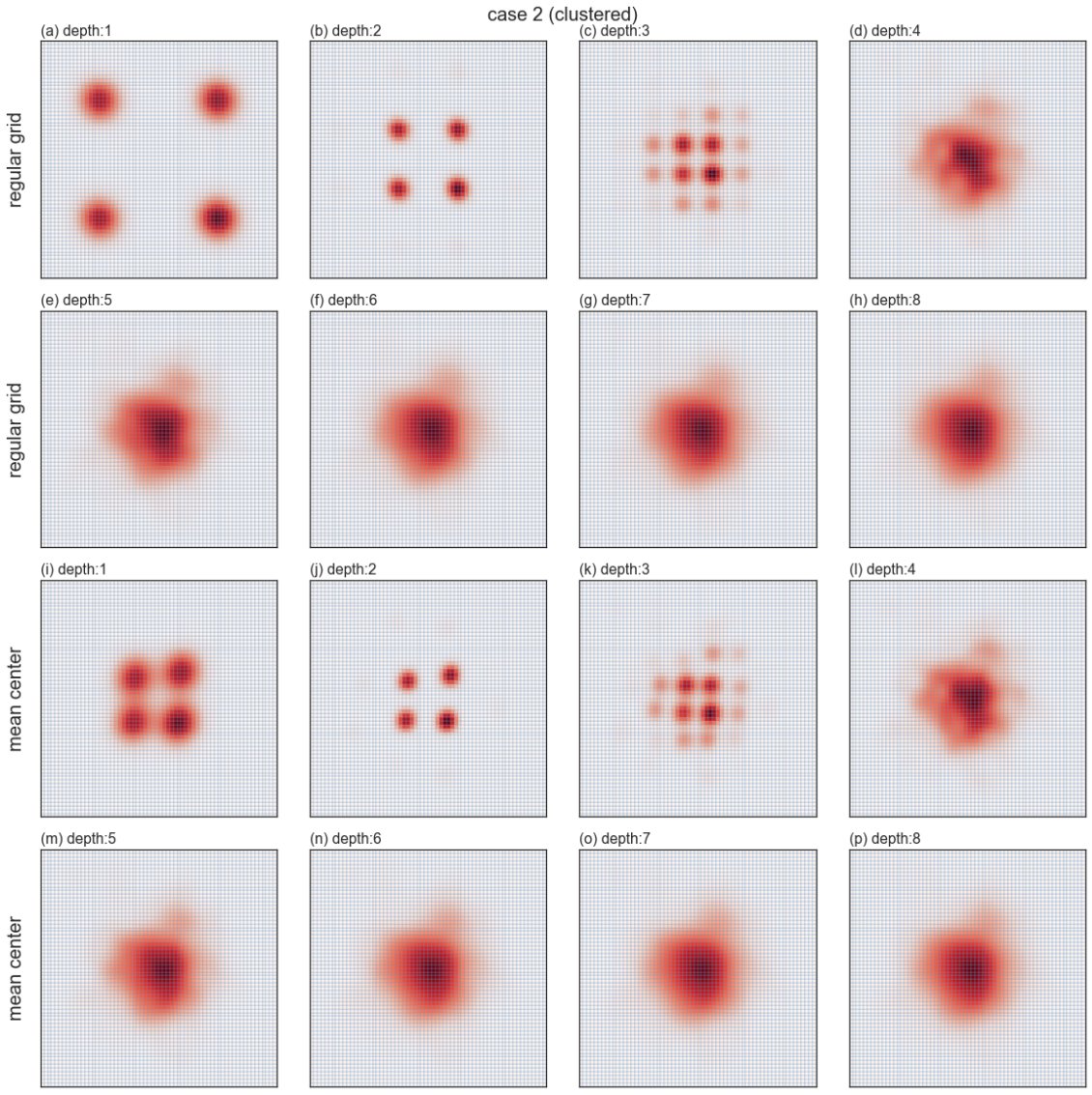


Figure 10: The Weighted KDE of case 2. The sub-figures in the first two rows (a-h) are calculated using grid center approach, whereas the sub-figures on the last two rows (i-p) are calculated using mean center approach. Each sub-figure within the rows indicates the aggregation using a depth, starting from one to eight.

doi:10.6342/NTU201802404

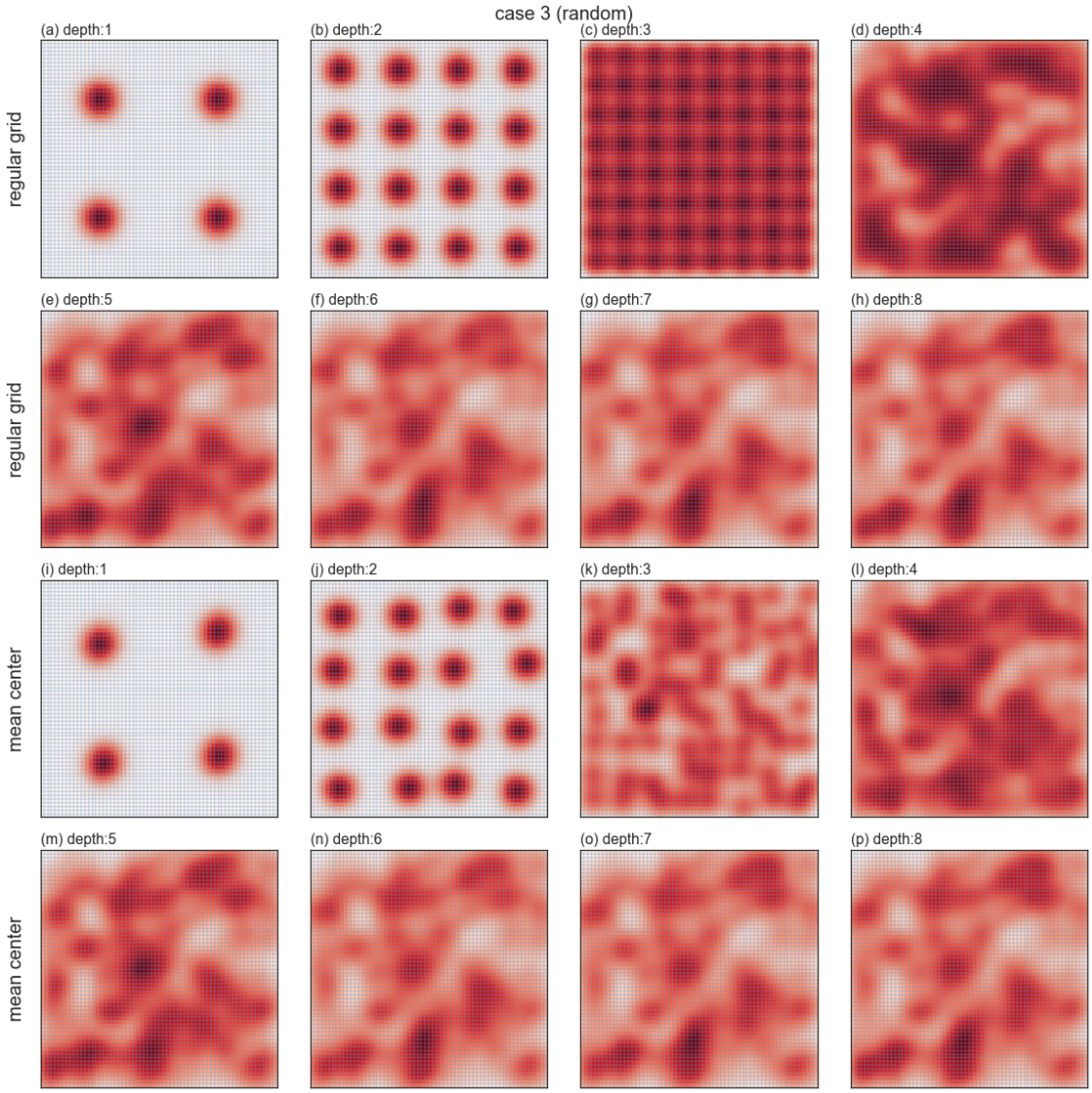


Figure 11: The KDE of case 3 without considering the counting weight. The sub-figures in the first two rows (a-h) are calculated using the grid center approach, whereas the sub-figures on the last two rows (i-p) are calculated using the mean center approach. Each sub-figure within the rows indicates the aggregation using a depth, starting from one to eight.

doi:10.6342/NTU201802404

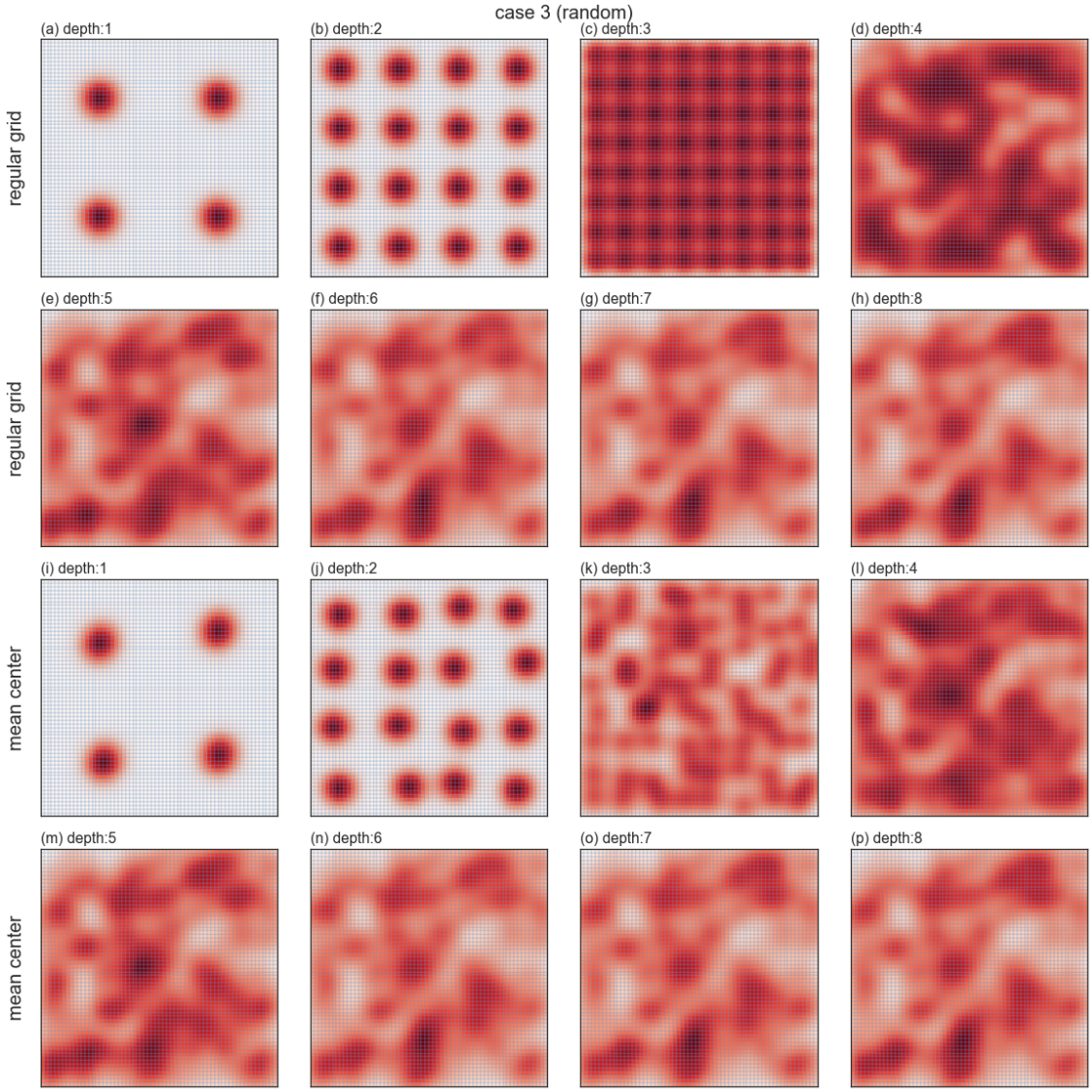


Figure 12: The Weighted KDE of case 3. The sub-figures in the first two rows (a-h) are calculated using grid center approach, whereas the sub-figures on the last two rows (i-p) are calculated using mean center approach. Each sub-figure within the rows indicates the aggregation using a depth, starting from one to eight.

doi:10.6342/NTU201802404

Appendix III: Extended analyses of experiment two

This appendix shows the extended analyses of experiment two, using each combination of parameter sets. The three parameters included the location of the cluster center, which is five values ranging from corner to the center of study area ((31.25,31.25), (62.5,62.5), (125,125), (250,250), (500,500)); the area size of cluster, which is a degree of separation in the random cluster model (as shown in Appendix I) that contain seven values (2^{-6}, 2^{-6}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}); the number of points, which is representing the density of cluster due to the fixed study area, contains eight values (2^{4}, 2^{5}, 2^{6}, 2^{7}, 2^{8}, 2^{9}, 2^{10}, 2^{11},). Each of the combinations was used to generate 99 sets of distributions for the test. In the following figures, the y-axis is used to represent the targeted index, the x-axis is used to show the changes of the cluster's center, varying sub-plots for showing the area size of a cluster, and the color of lines is used to shows the differences in cluster densities. In summary, these results show consistent findings as the main text of experiment two is showing.

doi:10.6342/NTU201802404

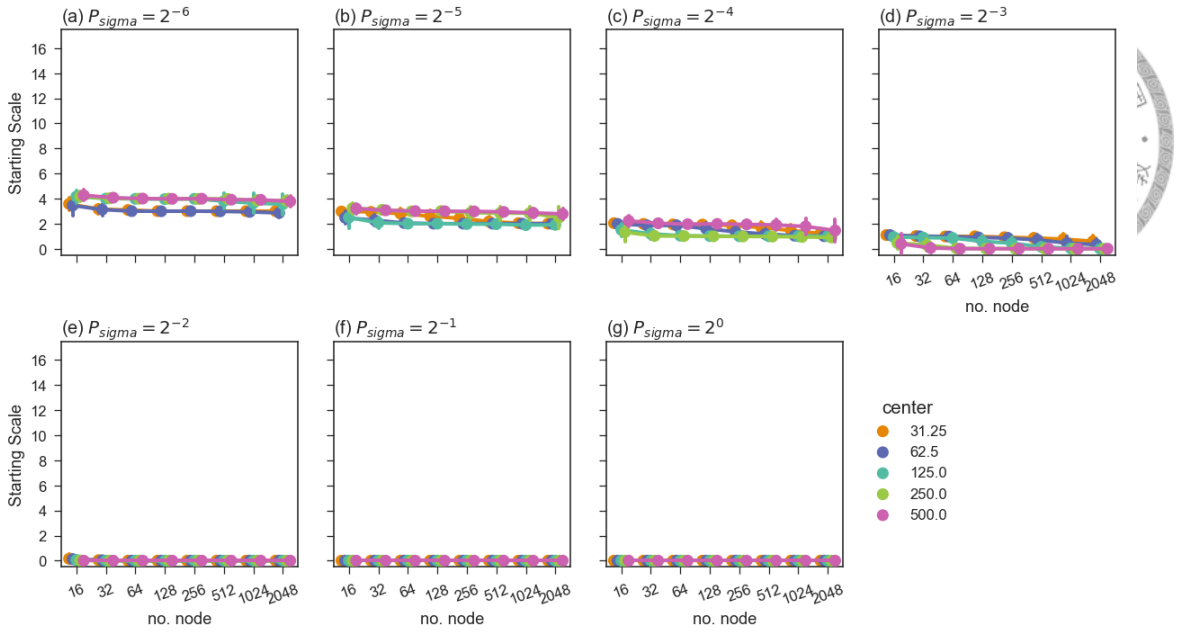


Figure 13: The changes of starting scale along with the three clustering parameters: (1) y-axis: starting scale; (2) x-axis: total number of points; (3) sub-plots: area size of cluster; (4) colors of lines: location of the cluster's center. The point markers indicate the average of the 99 sets of point distribution with the same parameter combinations, whereas the error bars indicate standard deviation.

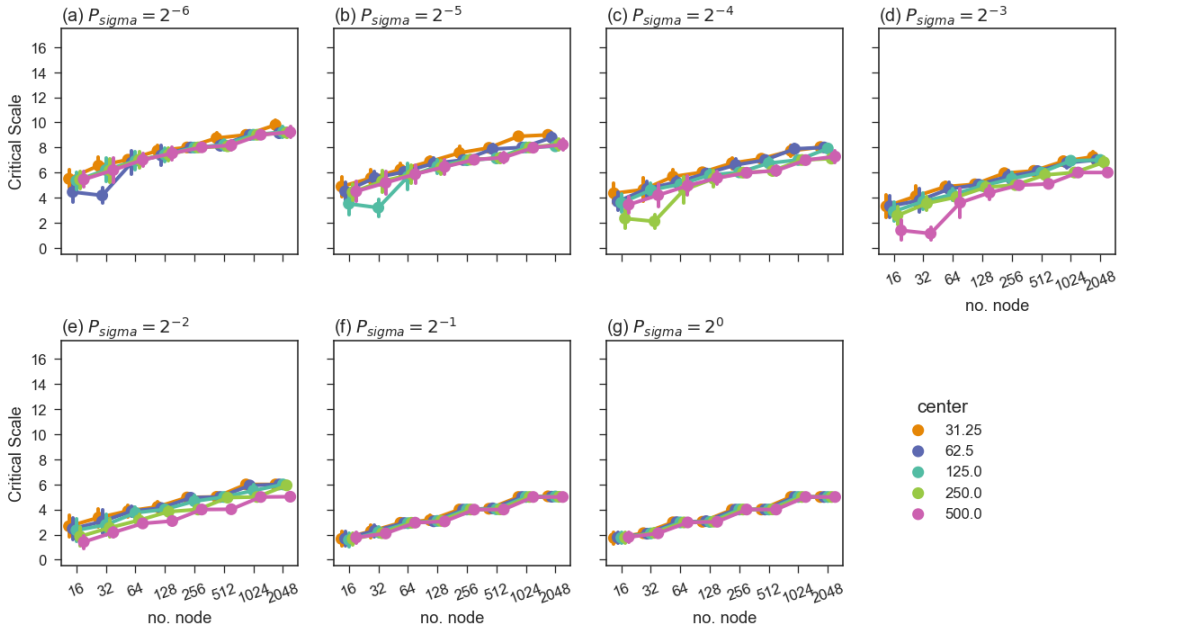


Figure 14: The changes of critical scale along with the three clustering parameters: (1) y-axis: critical scale; (2) x-axis: total number of points; (3) sub-plots: area size of cluster; (4) colors of lines: location of the cluster's center. The point markers indicate the average of the 99 sets of point distribution with the same parameter combinations, whereas the error bars indicate standard deviation.

doi:10.6342/NTU201802404

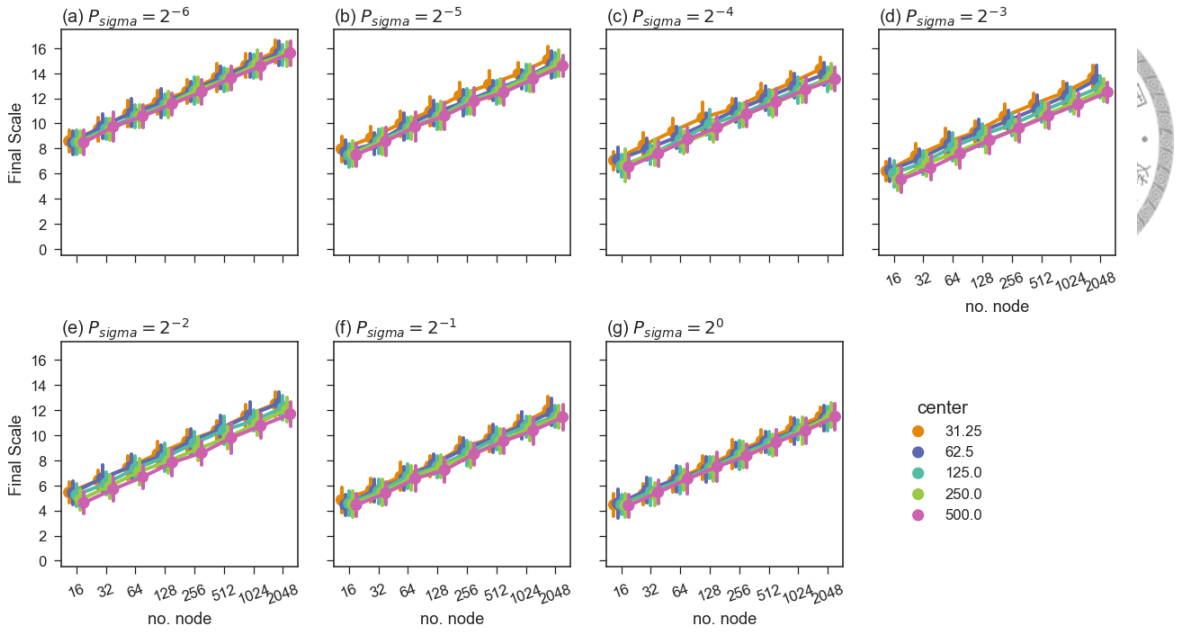


Figure 15: The changes of final scale along with the three clustering parameters: (1) y-axis: final scale; (2) x-axis: total number of points; (3) sub-plots: area size of cluster; (4) colors of lines: location of the cluster's center. The point markers indicate the average of the 99 sets of point distribution with the same parameter combinations, whereas the error bars indicate standard deviation.

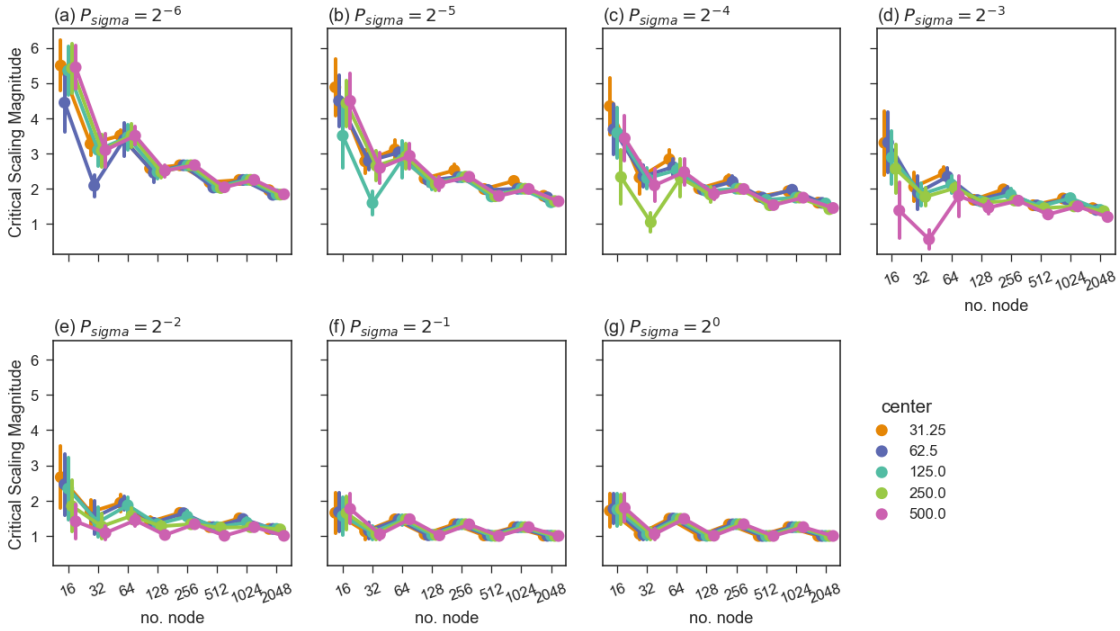


Figure 16: The changes of critical scaling magnitude along with the three clustering parameters: (1) y-axis: critical scaling magnitude; (2) x-axis: total number of points; (3) sub-plots: area size of cluster; (4) colors of lines: location of the cluster's center. The point markers indicate the average of the 99 sets of point distribution with the same parameter combinations, whereas the error bars indicate standard deviation.

doi:10.6342/NTU201802404

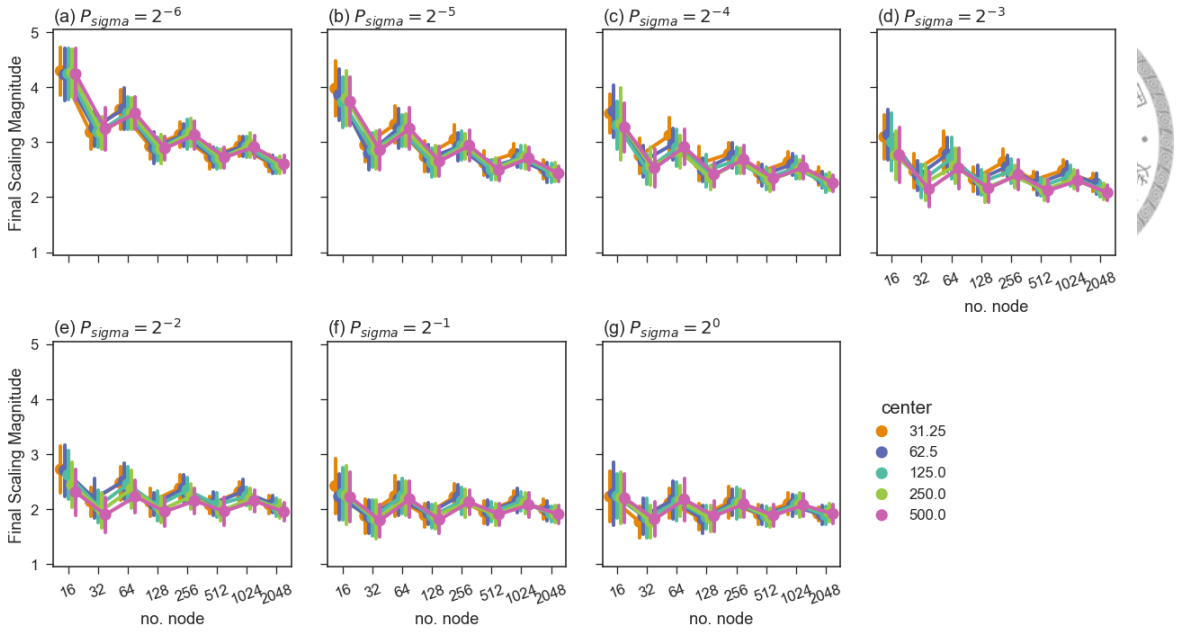


Figure 17: The changes of final scaling magnitude along with the three clustering parameters: (1) y-axis: final scaling magnitude; (2) x-axis: location of the cluster's center'; (3) sub-plots: area size of clusters; (4) colors of lines: the total number of points. The point markers indicate the average of the 99 sets of point distribution with the same parameter combinations, whereas the error bars indicate standard deviation.

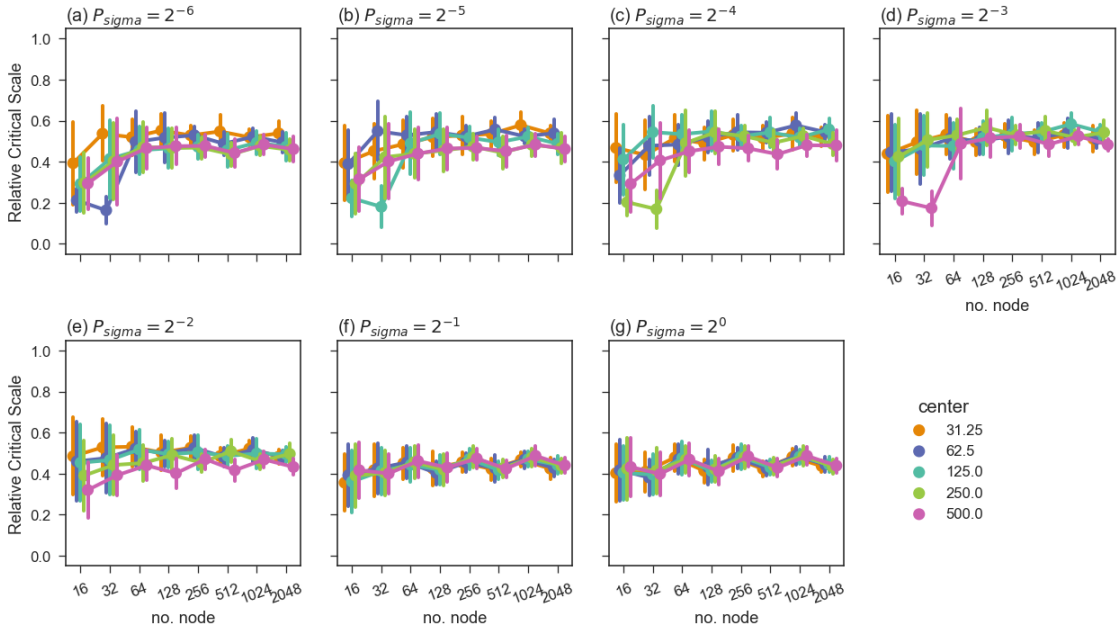


Figure 18: The changes of relative critical scale along with the three clustering parameters: (1) y-axis: relative critical scale; (2) x-axis: location of the cluster's center'; (3) sub-plots: area size of clusters; (4) colors of lines: the total number of points. The point markers indicate the average of the 99 sets of point distribution with the same parameter combinations, whereas the error bars indicate standard deviation.

doi:10.6342/NTU201802404

Appendix IV: Comparing grid center and mean center approach

To compare the RMSE results between the GC and MC approaches, the delta RMSE ($\Delta RMSE$) is calculated (Equation 3). The results is shown in Figure 19. The lower value of RMSE indicates the better result. While the $\Delta RMSE$ is calculated using RMSE of MC minus RMSE of GC, the positive value indicates the GC approach is better while negative value suggests that MC has a better result.

$$\Delta RMSE = \frac{RMSE_{MC} - RMSE_{GC}}{Max(RMSE_{GC}, RMSE_{MC})} \tag{3}$$

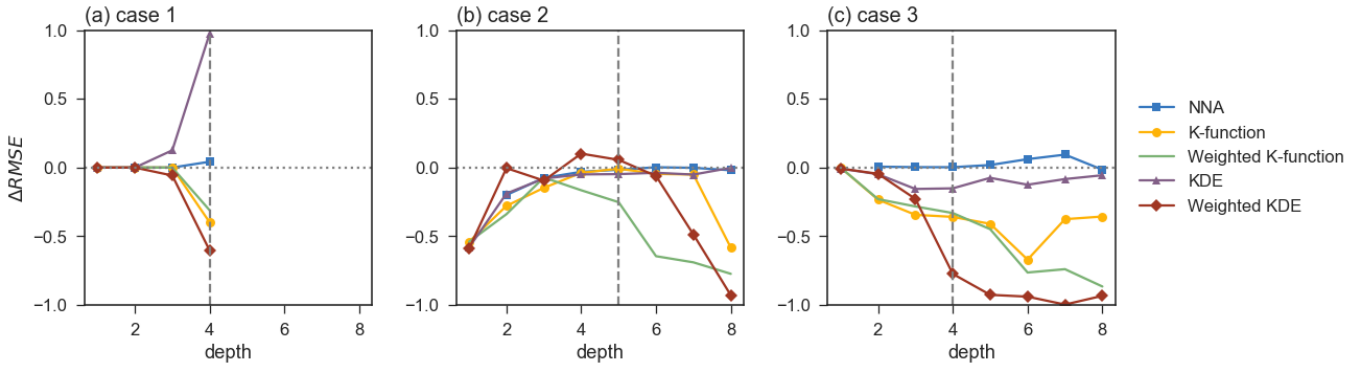


Figure 19: The differences of RMSE between the RC and MC approaches in experiment 1. The delta value above 0 suggests that the GC approach has a lower RMSE, and the delta value below 0 suggests that the MC approach has a lower RMSE. The delta value near to 0 indicates the differences is low, whereas the delta value equal to ± 1 means that one of the RMSE is equal to 0 while the other is not.

Based on Figure 19, all lines tend to be below the zero line (dotted grey horizontal line). These suggested that the MC approach is on average better than the GC approach. For case 2, the differences of RMSE on the critical scale (the dashed grey vertical line)

doi:10.6342/NTU201802404

are small (around zero). For case 3, the RMSE differences of NNA and KDE analyses are small; both of the K-function and weighted K-function have lower RMSE with MC, which suggested that the K-functions are more sensitive on the locations of aggregation points; the weighted KDE is near to -1.0, which means the RMSE of weighted KDE using MC is almost equal to zero on critical scale, while the same analysis using GC resulted in somewhere different from the result of original data.

In summary, focusing on the critical scale, the grid center approach of points aggregation is sufficient for capturing overall patterns of point distribution. Among the two approaches, the mean center approach will averagely result in better performance (i.e. lower differences between the spatial patterns of aggregation points and the original points). Therefore, the MC approach is suggested to be used while applicable. But, if the data set is large and the calculation time is a concern, the GC approach is an alternative that can produce sufficient analysis outcomes for capturing the spatial pattern.

doi:10.6342/NTU201802404

Appendix V: Five categories of empirical point distribution case studies

Aims

In the third experiment, this study illustrates the usage and analysis results using the perspective of point scaling using real-world data. The objective of this experiment is to understand the scaling properties of the point events in different categories at the same study area and to compare the spatial pattern of the categories in the scaling perspective.

The cases and datasets

The study area of this experiment is a square area that contains the whole Taiwan main island and also includes the nearby area. The data in this experiment is composed by five categories, including the locations of thunders, the epicenters of earthquakes, the observations of migratory birds, the patients of Dengue Fever cases, and the paths of flight. Three of the categories (thunder, earthquake, and dengue fever) are accessible from the Taiwan Open Data website (https://data.gov.tw/dataset/, dataset ID for thunder (9712), dengue fever (21025), earthquake (12730)). The bird observation is provided by the ebird database (https://ebird.org). The flight path data is provided by the ADS-B Exchange database (https://www.adsbexchange.com). To minimize the distortion for the study area, the coordinates of the point locations in all categories are projected into a modified projection, which is based on the Asia Lambert Conformal Conic projection (ESRI: 102012). Under this projection[1], the study area contain the area of x-coordinates between $1,300,000m$ and $1,900,000m$, and the y-coordinates between $2,700,000m$ and $3,300,000m$.

[1] +proj=lcc +lat_1=30 +lat_2=62 +lat_0=0 +lon_0=120 +x_0=1500000 +y_0=0 +datum=WGS84 +units=m +no_defs

doi:10.6342/NTU201802404

The five categories of data represented different distribution concepts. The occurrences of thunder do not directly restrict by the landscape; therefore, it spreads over the study area. The earthquakes happened within the study area located along the belt of fault. These two categories contained the natural events that were collected automatically by sensors and processed by professionals. The migratory birds' observation data represents the appearances of birds (a natural event), but the data is recorded manually by professionals or amateurs birdwatchers, thus the recorded point in this dataset is limited not only to the activity spaces of birds but also to the places where people can reach. In other words, the dataset represents the events of human-birds interaction. Dengue Fever is a vector-based human disease occurred in Taiwan that has an annual cycle (Chin et al., 2017[2]), with two types of patients (Wen et al., 2016[3]): indigenous cases that are infected within Taiwan; and imported cases that may be infected somewhere else according to their travel history, and become sick in Taiwan (Wen et al., 2016). Only the indigenous cases were included in this study. The dataset is recorded as the number of patients in each basic statistical areas (BSA, the finest spatial unit of statistical data from the government). This study randomly relocated the points into the area of the BSA for the analysis purpose. The flight path data is recorded through the Automatic Dependent Surveillances-Broadcast system and is representing the movement of flights. This dataset periodically recorded the location of each flight, thus the points in this dataset show the trajectories of flights, and is distributed along the air lanes. Therefore, this dataset is different from the other point events datasets aforementioned.

The datasets of the five categories were divided into cases for the analysis. For the thunder category, ten years of data (2004 - 2013) is used; the six months (April - September) – the six months with the largest number of thunder – is considered in the dataset, and each of the months is used as one case of this category; hence a total of 60 cases. For the earthquake category, the yearly recorded epicenters from 2008 to 2017 is used as a case; hence a total of 10 cases. For the birds category, 18 types of migratory birds that

[2]Chin W. C. B., Wen T. H., Sabel C. E. & Wang I. H. (2017). A geo-computational algorithm for exploring the structure of diffusion progression in time and space. Scientific Reports 7: 12565.

[3]Wen, T. H., Tsai, C. T., & Chin, W. C. B. (2016). Evaluating the role of disease importation in the spatiotemporal transmission of indigenous dengue outbreak. Applied Geography 76: 137-146.

doi:10.6342/NTU201802404

are observed to have appeared in Taiwan are used as the 18 cases, including *Accipiter soloensis, Anas zonorhyncha, Ardea alba, Ardea cinerea, Charadrius dubius, Chlidonias hybrida, Himantopus himantopus, Mesophoyx intermedia, Motacilla cinerea, Phalacrocorax carbo, Platalea minor, Pluvialis fulva, Recurvirostra avosetta, Tringa glareola, Tringa ochropus, Tringa stagnatilis, Upupa epops, Vanellus vanellus*. The Dengue Fever category included the 18 years (starting from 1998 till 2015) of data, which are transformed to 18 cases; each of the 18 cases contains the events happened from April 1st of the year till the March 31 of the following year, due to the annual cycle of Dengue Fever in Taiwan. Finally, the flight path data included four days (2018/4/14 - 2018/4/17, two days of the weekend and two days of weekday) of data, and is separated hourly; hence a total of 96 cases.

The number of cases and the number of points in each category is shown in Table 1 and the box plot of the case sizes of the five categories is shown in Figure 20. The magnitudes of the case size (number of points in each case) varied among the different categories. The case sizes of thunder category, which median is about 292 thousands points, is far larger than the other categories. The earthquake category has the smallest variation between cases, which also has the lowest number of total points in all of the cases. The case size of Dengue Fever has the second largest variation among its cases, which included three outliers.

Table 1: The number of points in each cases and category.

Category	total point	case number	smallest case	largest case	case size median (M.A.D.)
Thunder	22305355	60	8174	1370670	292431.5 (244469.967)
Earthquake	6257	10	465	769	639.0 (84.300)
Birds	102967	18	1051	14772	4250.5 (3456.698)
Dengue Fever	74636	18	24	43327	739.0 (5749.519)
Flight path	321643	96	980	5302	3719.0 (1151.582)

Figure 21 shows the example of point spatial distribution in each category. All of the point spatial distribution are included in the supplementary. As shown in Figure 21a, which is an average case, the point locations of thunder spread over most of the study area. Earthquake on the other hand (Figure 21b), shown several clusters along the east

doi:10.6342/NTU201802404

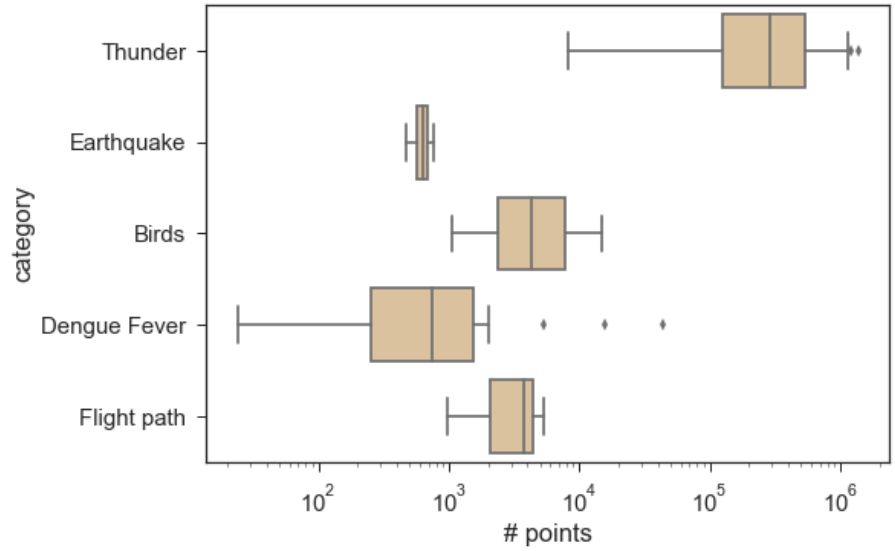


Figure 20: The case size (number of points in each case) distribution.

coast of Taiwan, and several on Taiwan. The birds' observations showed more clusters over the land area (Figure 21c), which is mainly located in the plain area and coastal area of Taiwan. The Dengue Fever (Figure 21d) showed even more clustered patterns at the southern part of Taiwan. The flight path (Figure 21e) showed several lines of air lanes mainly flying on both sides of the Taiwan Strait.

doi:10.6342/NTU201802404

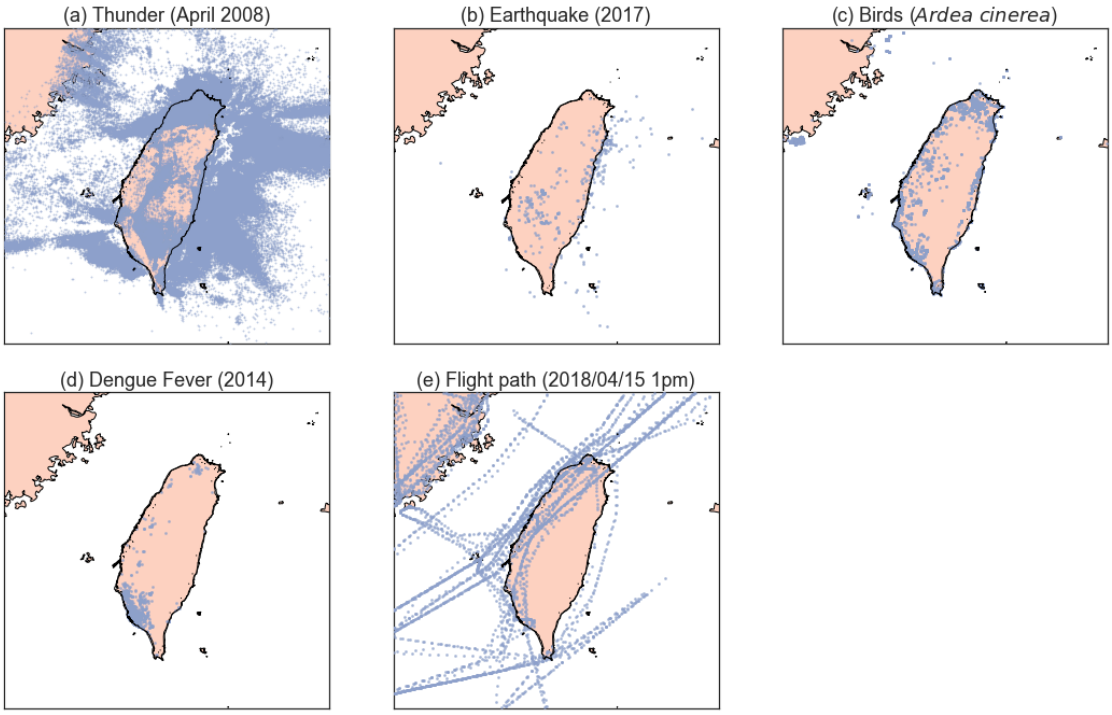


Figure 21: The examples of points' spatial distribution of a case in each category: (a) the thunder distribution in April 2008; (b) the earthquake epicenters in 2017; (c) the observations of *Ardea cinerea; (d) the distribution of patients in 2014; (e) the flight path during 13:00-13:59 on April 15th, 2018.*

doi:10.6342/NTU201802404

Results

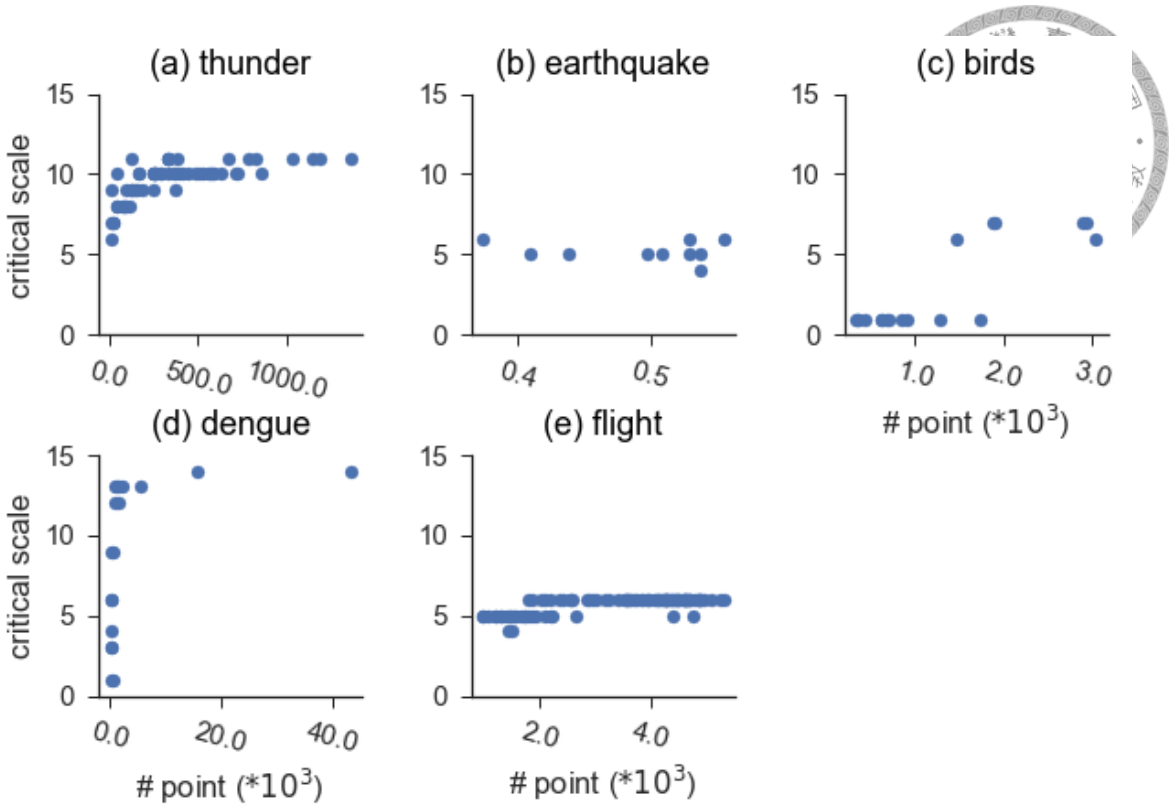


Figure 22: The critical scale against the number of point in the cases (thousands) of the five categories.

doi:10.6342/NTU201802404

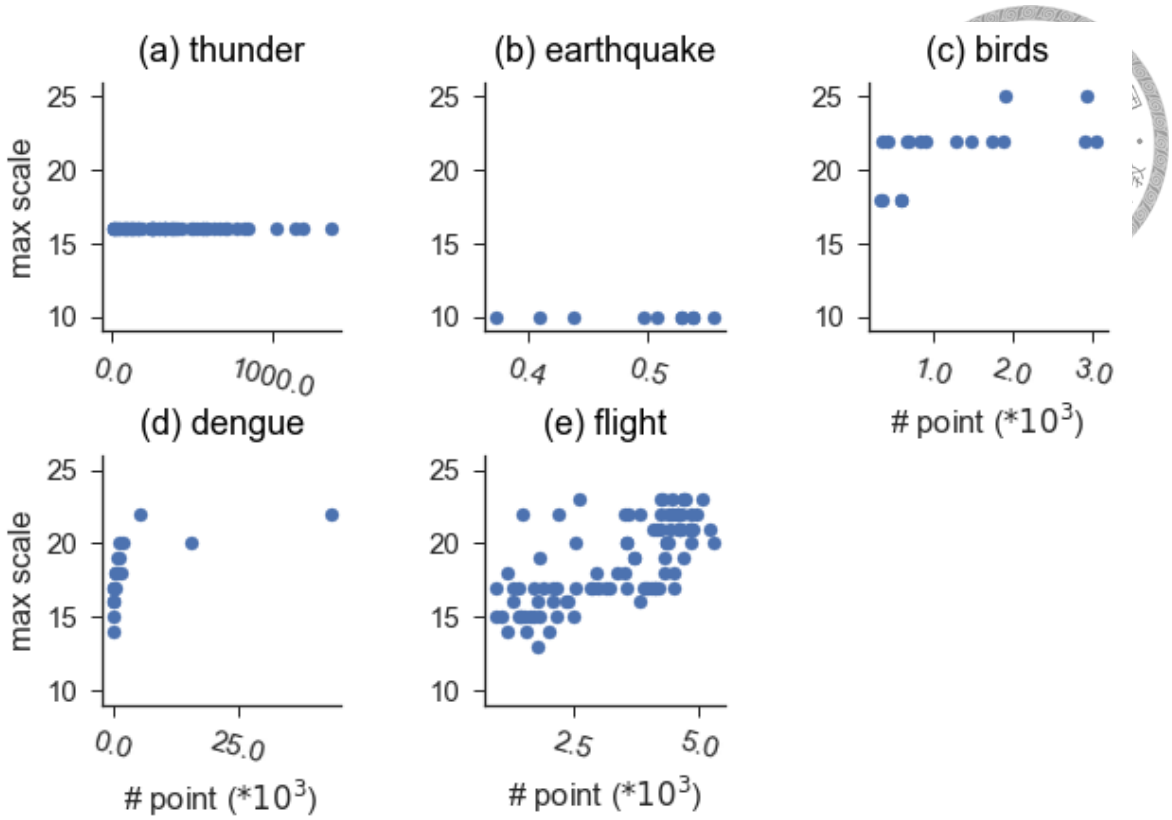


Figure 23: The max scale (final scale) against the number of points in the cases (thousands) of the five categories.

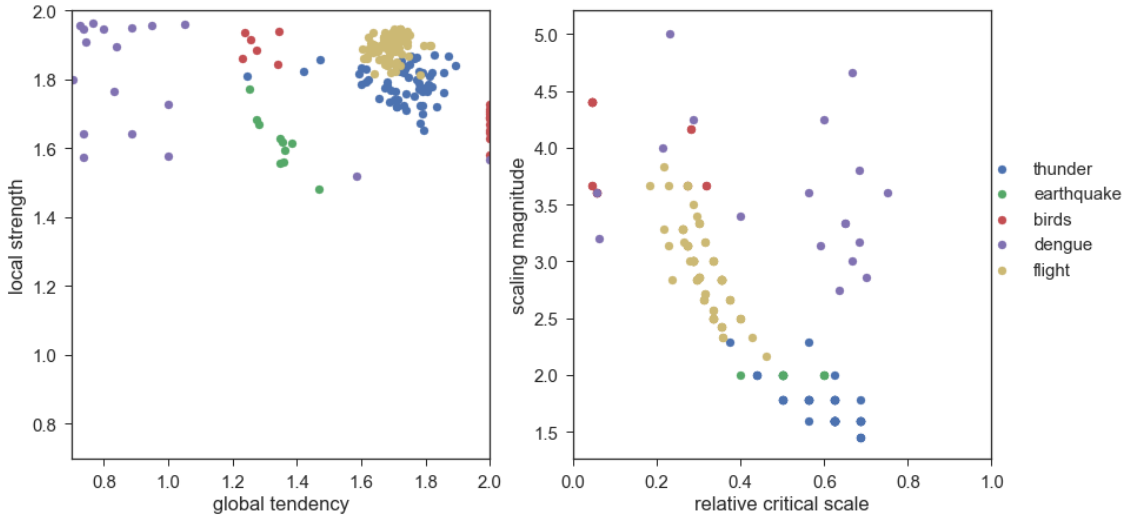


Figure 24: The measurements results of the five categories: (a) Local strength (local fractal dimension, Y-axis) vs. global tendency (global fractal dimension, X-axis) of scaling, (b) the relative critical scale (X-axis) vs. scaling magnitude (final scaling magnitude, Y-axis)

doi:10.6342/NTU201802404