

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Estatística

Vinicius Ricardo Riffel

**A Stochastic Approach to Establish a Metric to Quantify the Modifiable
Areal Unit Problem**

Belo Horizonte
2024

Vinicius Ricardo Riffel

**A Stochastic Approach to Establish a Metric to Quantify the Modifiable
Areal Unit Problem**

Final Version

Thesis presented to the Graduate Program in Statistics of the
Federal University of Minas Gerais in partial fulfillment of the
requirements for the degree of Master in Statistics.

Advisor: Renato Martins Assunção

Belo Horizonte
2024

Riffel, Vinicius Ricardo.

R564s

A Stochastic approach to establish a metric to quantify to modifiable areal unit problem [recurso eletrônico] / Vinicius Ricardo Riffel – 2024.

1 recurso online (39 f. il., color.): pdf.

Orientador: Renato Martins Assunção.

Dissertação (Mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.

Referências: f. 36-39.

1. Estatística – Teses. 2. Análise espacial (Estatística) – Teses. 3. Análise por conglomerados – Teses. 3. Análise estocástica. I. Assunção, Renato Martins. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2(043)



UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA




FOLHA DE APROVAÇÃO

A Stochastic Approach to Establish a Metric to Quantify the Modifiable Areal Unit Problem.


VINICIUS RICARDO RIFFEL

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTATÍSTICA, como requisito para obtenção do grau de Mestre em ESTATÍSTICA, área de concentração ESTATÍSTICA E PROBABILIDADE.


Aprovada em 27 de junho de 2024, pela banca constituída pelos membros:

Documento assinado digitalmente
 **RENATO MARTINS ASSUNCAO**
Data: 04/07/2024 18:21:54-0300
Verifique em <https://validar.iti.gov.br>

Prof. Renato Martins Assunção - Orientador
DCC/UFMG

Documento assinado digitalmente
 **ROSANGELA HELENA LOSCHI**
Data: 28/06/2024 08:30:09-0300
Verifique em <https://validar.iti.gov.br>

Profª. Rosangela Helena Loschi
DEST/UFMG

Documento assinado digitalmente
 **PAULO JUSTINIANO RIBEIRO JUNIOR**
Data: 05/07/2024 15:54:23-0300
Verifique em <https://validar.iti.gov.br>

Prof. Paulo Justiniano Ribeiro Junior
DEST/UFPR

**Kevin A.
Butler**

 Digitally signed by Kevin A. Butler
DN: cn=Kevin A. Butler, o=Esri, ou,
email=kbutler@esri.com, c=US
Date: 2024.07.10 07:27:46 -07'00'

Prof. Kevin Butler
ESRI INC.

Belo Horizonte, 27 de junho de 2024.

Acknowledgments

The author was financially supported by FAPEMIG, CNPQ, CAPES and FCO.

Resumo

O Problema da Unidade Areal Modificável (MAUP) afeta significativamente os resultados da análise espacial ao variar os resultados com base na escala e zonificação das unidades geográficas utilizadas. Esta tese introduz uma abordagem estocástica inédita para quantificar os efeitos do MAUP, apresentando um índice que mede a sensibilidade das análises espaciais às mudanças nas configurações das unidades areais. A metodologia proposta é baseada no algoritmo SKATER e pode ser utilizada em qualquer análise espacial. Aplicamos o método proposto a cerca de 2.000 diferentes conjuntos de dados. Os resultados indicam que os impactos mais pronunciados do MAUP ocorrem em escalas menores, onde a agregação das áreas altera significativamente os resultados estatísticos. O estudo também revela uma alta correlação entre os efeitos de escala e zonificação, sugerindo a natureza interligada desses componentes. Embora os índices propostos forneçam uma ferramenta valiosa para avaliar o MAUP, desafios computacionais em grandes conjuntos de dados destacam a necessidade de otimizações algorítmicas adicionais.

Palavras-chave: problema da unidade areal modificável; construção de zonas automatizadas; aprendizado não supervisionado; dependência de zonificação; partição espacial; clusters espaciais.

Abstract

The Modifiable Areal Unit Problem (MAUP) significantly affects spatial analysis outcomes by varying results based on the scale and zoning of the geographical units used. This thesis introduces a novel stochastic approach to quantify the MAUP effects, presenting an index that measures the sensitivity of spatial analyses to changes in areal unit configurations. The proposed methodology is based on the SKATER algorithm and can be used in any spatial analysis. We applied the proposed method to around 2,000 different datasets. The findings indicate that the most pronounced impacts of MAUP occur at smaller scales, where area aggregation significantly alters statistical outcomes. The study also reveals a high correlation between scale and zoning effects, suggesting the intertwined nature of these components. While the proposed indices provide a valuable tool for evaluating MAUP, computational challenges in large datasets highlight the need for further algorithmic optimizations.

Keywords: modifiable areal unit problem; automated zonation construction; unsupervised learning; zonation dependence; spatial partitioning; spatial clusters.

List of Figures

1.1	The numbers in the upper corner of each sub-image represent the count of points in each group. The title of each subfigure indicates the variance in the total number of points across groups. Subfigure (a) displays the data at the individual level. The zoning effect is illustrated by comparing the second and third plots. The scale effect is demonstrated by comparing the second and fourth plots.	12
1.2	Maximum and minimum curves for calculating the Pearson correlation between household yearly income and commuting time (left) and number of rooms (right). The maximum s scale corresponds to the Census tract level for Oakland County, MI (left) and Riverside, CA (right). Data are from the American Community Survey (ACS), year 2020.	13
2.1	Left: Map of an urban area transformed into a graph $\mathcal{G} = (V, E)$. Right: A spatial partitioning $z = (\mathcal{G}_1, \dots, \mathcal{G}_4)$ of the graph with 4 zones or spatial clusters.	18
2.2	Left: Two spanning trees associated with the adjacency graph $\mathcal{G} = (V, E)$ shown in Figure 2.1. Third plot: Partitioning of the graph in 5 spatial clusters by pruning 4 edges from the spanning tree. The pruned edges are shown in red. Right: Spatial clusters determined by the pruning of the red edges in the previous spanning tree.	20
2.3	Comparing single-shot estimator (dotted line) with the estimates of M and m (solid red lines). Each black line represents an output of the algorithm 2.3 using the usual Pearson correlation coefficient as $T(s, z)$. Data is from Riverside County of 2020 ACS. The correlation was evaluated between the proportion of the black population and household income.	24
3.1	Map of Pennsylvania Divided into ACS Census Tracts	28
3.2	Boxplot of each variable across the states.	29
3.3	Typical behaviour of $M(s)$ and $m(s)$. The solid blue line indicates the correlation at the original scale. Upper left: Harris County (TX) using household income and proportion of black population. Upper right: Franklin County (OH) using proportion of the black population and number of rooms. Lower left: San Joaquin County (CA) using household income and proportion of black population. Lower right: Washtenaw County (MI) using commuting time and household income.	30

3.4	Boxplot of I_z for each pair of variables and state.	31
3.5	Plot of I_z for each county and pair of variables against its respective correlation at the original scale. The blue line is the estimates using the LOESS method.	32
3.6	Cumulative effect of I_z on the respective percentiles at the original scale. The black line is the percentile average of accumulated I_z . The red lines indicate the 5th and 95th of the accumulated I_z	32
3.7	Typical behaviour on the random scenario. Each black line represents a simulation from $\mathcal{L}(s)$. The solid blue line indicates the correlation at the original scale. The solid red line indicates the maximum and minimum at each scale. Upper left: Wilson County (NC) using commuting time and proportion of black population. Upper right: DuPage County (IL). Lower left: Lee County (FL). Lower right: Union County (NC), the latter three the correlation was calculated between commuting time and household income.	33
3.8	Boxplot of I_m for each pair of variables. HI stands for household income, BP for the proportion of the Black population, NR for the number of rooms and Ct for commuting time.	34
3.9	Plot of both MAUP measurements. The solid line denotes the identity line.	34
3.10	Plot of the scale effect (I_s) and the zoning effect (I_z). Both of measurements are highly correlated.	34

List of Algorithms

2.1	SKATER	21
2.2	Obtaining random partitions	21
2.3	Estimating $M(s)$ when $T(s, z)$ is the correlation	23
2.4	Estimating $M(s)$ in an efficient way when $T(s, z)$ is the correlation	24

Contents

1	Introduction	11
1.1	Related work	15
2	Methodology	17
2.1	Overview of our method	17
2.1.1	Definitions and notation	17
2.1.2	The SKATER algorithm	19
2.2	Estimating $\mathcal{L}(s)$ and the extreme bounds $M(s)$, and $m(s)$	21
2.3	The MAUP index	25
2.3.1	The scale effect	26
3	Results	28
3.1	Empirical analysis	28
3.1.1	Datasets	29
3.1.2	Results	30
3.1.3	Random scenario	32
4	Conclusions	35
	References	36

Chapter 1

Introduction

The analysis of geographical data often requires partitioning the continuous space into discrete units, as seen in examples such as disease mapping [18, 32, 17, 28], electoral redistricting [14, 13, 26], and image analysis [3, 6, 16].

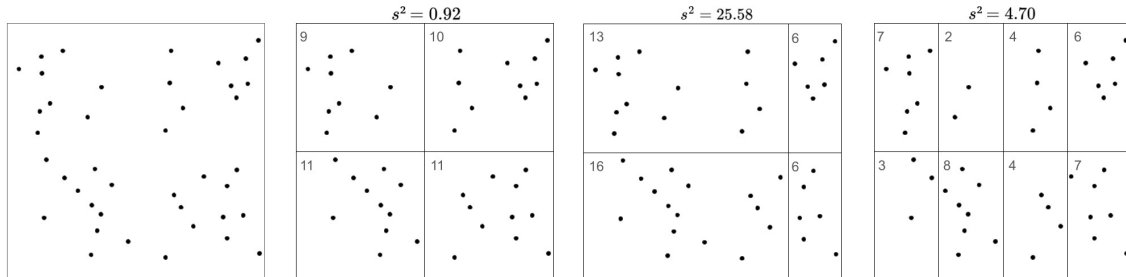
In the case of disease mapping, partitioning is necessary for reasons of confidentiality (preventing the exact address locations of disease cases from being disclosed) or data compatibility (aligning population, economic factors, and disease counts). For electoral redistricting, partitioning ensures compliance with legislation mandating specific electoral district divisions to elect representatives. In image analysis, technological constraints require averaging sensor measurements within a given pixel.

The prevalence of geographical data analysis relying on discrete spatial units raises questions about the impact of partitioning on the analysis' conclusions. Analysts have some discretion in determining the number, shape, and size of these spatial units. However, sometimes the geographical partitioning of the region is out of the control of the analyst, such as when she needs to use data from the Census Bureau's pre-established zoning. Regardless of whether the partitioning is chosen by the analyst or predetermined, it was recognized early on that this can significantly affect results [12, 35, 22, 23, 10]

Figure 1.1 illustrates this issue, showing how the variance in the number of points within each cell changes markedly with different forms of aggregation. Consider initially the second and third plots, each divided into four areas. The variance s^2 dramatically shifts from 0.92 in the second plot to 25.58 in the third plot. This substantial difference arises solely from the aggregation method—a phenomenon known as the *zoning effect* [23]. Now, compare the second and fourth plots, which have the same grouping shape (rectangles of equal sizes), but more groups in the fourth plot. Again, there is a five-fold variation induced by the scale at which the data is aggregated, termed the *scale effect* [23].

The potential effect of spatial partitioning in the data analysis is called the Modifiable Areal Unit Problem (MAUP). It has been a topic of study in the literature for decades, having first been explicitly identified in the 20th century by [12]. In 1950, [35] showed that the correlation between wheat yields and potato yield variables measured in counties of England could vary from 0.22 to 0.99 depending on how the spatial units

Figure 1.1: The numbers in the upper corner of each sub-image represent the count of points in each group. The title of each subfigure indicates the variance in the total number of points across groups. Subfigure (a) displays the data at the individual level. The zoning effect is illustrated by comparing the second and third plots. The scale effect is demonstrated by comparing the second and fourth plots.



Source: Developed by the author.

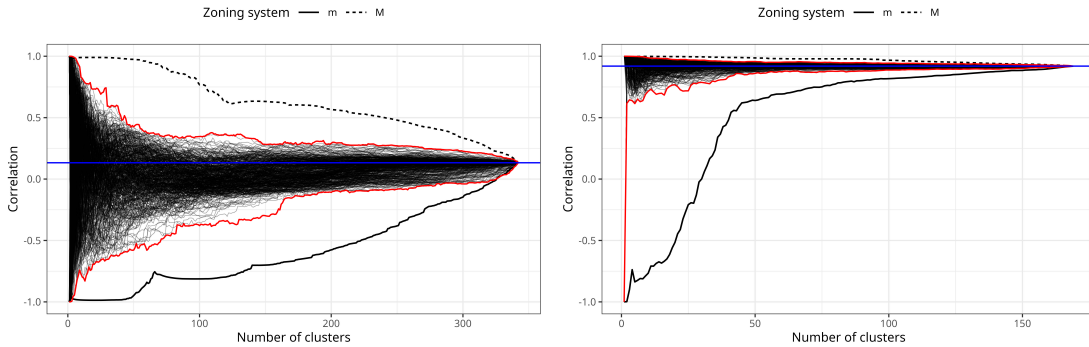
were aggregated. Later, MAUP was brought into prominence through the work of [22, 23]. They showed that one can provide two different geographical partitionings that ultimately lead to widely different conclusions in the data analysis. At the end of the 20th century, [10] analyzed the MAUP on the coefficients of a multivariate regression using US census data. They showed that a coefficient could vary from around 10,000 up to 18,000 depending only on how the spatial units were aggregated. They concluded that "(...) the modifiable areal unit problem is shown to be essentially unpredictable in its intensity and effects in multivariate statistical analysis and is therefore a much greater problem than in univariate or bivariate analysis. The results of this analysis are rather depressing in that they provide strong evidence of the unreliability of any multivariate analysis undertaken with data from areal units."

Can one obtain any desired measurement simply by organizing spatial data differently? In principle, yes, but this is not always feasible. Empirically, it has been observed that extreme sensitivity to the MAUP is not a universal characteristic [10, 30, 8]. In some data analyses, even a malicious adversarial partitioning approach fails to generate significant variation in results and measurements extracted from a given dataset.

Given that MAUP potentially impacts data analysis conclusions, it is crucial to assess its influence in each specific data analysis. Our motivation in this work is to establish a metric for quantifying the sensitivity of a given data analysis to the MAUP. With such a metric, one can determine if the MAUP is relevant to the analysis being conducted. If it is not, there is little concern about the specific partitioning adopted. However, if it is relevant, extra effort should be dedicated to quantitatively evaluating the sensitivity and variability of conclusions under different partitioning schemes. Additionally, justification for the selected partitioning becomes important.

Let $T(s, z)$ be the statistic of interest evaluated in a particular map partitioned into polygonal regions, such as the correlation between X and Y , two variables measured in each region. The parameter s denotes the scale or number of regions in which the

Figure 1.2: Maximum and minimum curves for calculating the Pearson correlation between household yearly income and commuting time (left) and number of rooms (right). The maximum s scale corresponds to the Census tract level for Oakland County, MI (left) and Riverside, CA (right). Data are from the American Community Survey (ACS), year 2020.



Source: Developed by the author.

map has been partitioned, while z serves as an index for the specific zoning adopted with the s regions. For each number s of regions into which the map may be partitioned, we estimate the probability distribution of $T(s, z)$ under all possible partitionings z .

We also obtain the $m(s)$ minimum and $M(s)$ maximum values that $T(s, z)$ can attain for each value of s . This establishes an envelope of viable outputs for $T(s, z)$ at any given scale s , as depicted in Figure 1.2. In both plots, the vertical axis represents the Pearson correlation coefficient between two variables, X and Y , at different scale levels s shown on the horizontal axis. The plot on the left utilizes data from the 2020 American Community Survey (ACS) at the Census tract level for Oakland County, Michigan, while the plot on the right pertains to Will County, Illinois. The variables considered are commuting time and yearly income on the left-hand side plot and the number of rooms and yearly income on the right-hand side plot. We averaged the variables after creating spatial clusters. For instance, transitioning from scale s to $s + 1$ involves creating a new group. The values for variables X and Y in this new group will be the averages of the units being grouped. Additionally, it is worth emphasizing that the correlation was obtained between the spatial units. For example, if a certain county has S Census tracts, the correlation was calculated by considering each tract as an individual observation.

The correlation at the largest s scale value represents the correlation at the finest resolution possible, which is the Census tract level. The horizontal blue line denotes this reference correlation value. As we aggregate adjacent areas, resulting in a smaller number of larger regions, the correlation between X and Y fluctuates within the bounds delineated by the curves $M(s)$ and $m(s)$.

Within the $M(s)$ and $m(s)$ bound curves, we see a random sample of trajectories. Each trajectory is obtained by sequentially creating random partitionings as we change the s scale. Assume that the finest resolution is that with $s = n$ regions when we use all

the n Census tracts. We randomly aggregate two regions obtaining $s = n - 1$ regions and a value for $T(n - 1, z)$. Next, we randomly aggregate two of the previous regions ending with $s = n - 2$ regions, a new partitioning z , and a corresponding value $T(n - 2, z)$. We keep sequentially aggregating two of the available regions randomly obtaining a new scale value, a coarser partitioning, and the corresponding $T(s, z)$. Each trajectory shown in Figure 1.2 represents one realization of this random sequential procedure. In some cases, as in the left-hand side plot, the curves $m(s)$ and $M(s)$ may represent extreme cases, that are not easily approached by the random trajectories.

Our main metric is based on the area determined by the envelope divided by its maximum possible value. We compare the total area enclosed by $M(s)$ and $m(s)$ with the maximum possible area. The MAUP index is defined as the ratio between the former area and the latter, jointly capturing the zoning and scale effects. A small ratio indicates low sensitivity to the MAUP, whereas a large ratio indicates extreme sensitivity to arbitrary partitioning. In the left-hand side plot, the correlation varies considerably around the reference blue line, even for scale levels s only slightly smaller than the original resolution. Conversely, in the right-hand side plot, the correlation remains relatively stable for most scale levels s , except in cases of unrealistic partitioning using fewer than 25 regions. The MAUP poses a significant challenge in the first analysis but has minimal impact in the second analysis.

This first metric represents a pessimistic view. The analyst assumes that the conclusions must be robust against all arbitrary partitionings. In particular, they consider an adversarial approach. At each scale (s), the analyst asks what partitionings would move the statistic under study furthest—both higher and lower—from the value it has at the most refined resolution. These partitionings could be the choices of an adversary aiming to interfere with the data analysis at the finest resolution, and they represent the maximum impact a partitioning can provoke at scale s in the study.

We consider a second metric for the MAUP effect, one less pessimistic. Rather than considering the maximum impact one could impose in an analysis at any scale s , we take the average maximum impact that can be made at each scale s . Select a random trajectory and consider its value $T(s, z)$ at scale level s with one specific partitioning z . Then, measure the maximum change we can attain at scale $s - 1$. Finally, average over all possible partitionings at scale s . This again is divided by its maximum possible value to obtain a value between 0 and 1.

Our new metrics to evaluate the robustness of spatial data analysis to the MAUP relies on a spatial partitioning method known as SKATER, introduced by [1]. This is an algorithm to partition a map into disjoint regions by aggregating connected areas. It optimizes the creation of homogeneous regions and has great speed [2]. This algorithm is crucial to make the calculations feasible.

Our method is general and does not presuppose any specific characteristics of

the T statistic under analysis. In this paper, we demonstrate how our method can be applied to various forms of $T(s, z)$: any univariate measure (such as mean, variance, or an inequality index like the Gini index), correlation coefficient between Y_1 and Y_2 , or regression coefficient β derived from a regression involving Y and a feature vector \mathbf{x} .

Section 1.1 provides an overview of how the MAUP problem has been addressed in the literature. We start our technical presentation in Section 2.1 with the definitions and notations. As the SKATER algorithm is central to our method, we present its original formulation. In Section 2.2, we discuss the process of obtaining the two enveloping curves $M(s)$ and $m(s)$ using SKATER. Additionally, we describe how the full probability distribution of T can be estimated for each scale level x . Subsequently, our two new metrics for quantifying the MAUP in data analysis are derived in Section 2.3. In Section 3.1, we run an extensive empirical analysis of the metrics. We close the paper with conclusions in Section 4.

1.1 Related work

MAUP has been a subject of discussion for decades. Perhaps the first to identify a MAUP effect were [12], where the authors presented illustrative examples of the scale effect. In particular, they showed a significant change in the correlation coefficient using multiple datasets. Another early study on the impacts of MAUP in the spatial analysis was done by [35]. The authors also analyzed the behaviour of the correlation coefficient in different aggregations of counties in England, also showing that it could vary widely depending on how the spatial unit were aggregated. The work of [12] was later replicated by [22], who looked at the zonation effect by maintaining a fixed scale and generating various map arrangements. The additional simulations showed that the correlation coefficient could vary even more than before, going from approximately -1 up to just below 1 only by rearranging the spatial units. These studies raised concerns about the methodology of spatial data analysis, demonstrating that the correlation coefficient could exhibit a wide variation depending on the chosen scale and zonation. [10] further underscored the potential impact of the MAUP on multivariate spatial analysis. They used data from the US and England censuses to show that the statistics of a multivariate regression (such as regression coefficients, R^2 , etc) were also vulnerable to the MAUP. Additionally, they showed that Moran's I could vary substantially due to MAUP, but the results with univariate statistics were less concerning than the results with multivariate statistics.

However, skepticism persists among some authors regarding the significance of the

MAUP in spatial analysis. [21], for instance, highlights the absence of practical examples demonstrating the costly consequences of disregarding the MAUP.

Whether considered a major problem or not, the MAUP often receives insufficient attention in spatial analysis. [20] identifies two primary reasons for this oversight: firstly, a lack of engagement with the creation of areal units due to the prevalence of pre-aggregated spatial data, and secondly, an assumption that the MAUP does not significantly impact the analysis. Recent observations, as highlighted by [27], suggest that this situation persists, with even top-tier journals frequently neglecting to address the MAUP, particularly its zonation effect. Even though, MAUP has been the subject of numerous empirical investigations [24, 10], particularly in the context of census data, where arbitrary scales defined by the census authorities are often utilized. [19] explored its implications on ecological models for air pollution, evaluating differences in coefficients across multiple scales. Recent studies have also addressed the MAUP in various applications. For instance, [34] examined the MAUP's impact on omission errors (when independent variables are missing from a regression model). [33, 11, 36, 15] analyzed the MAUP in the context of urban mobility. They showed that MAUP may play a key role in this field as well. In geographical health, [28] proposed a method for alleviating the MAUP effect on single disease maps, and [30] applied this methodology to mental health emergency data.

Various methodologies have been developed to quantify the MAUP. To assess the MAUP effect, in general, the authors compute a statistic in several map arrangements in different scales. For instance, [4] proposed a Bayesian share-effect model, which considers the MAUP as a parameter within a Bayesian regression framework. This approach allows for the assessment of the MAUP effect both locally and globally, providing interpretable scales and credible intervals. The MAUP parameter facilitates comparisons among maps, however, it lacks an upper limit. [29] addressed the MAUP by simulations of different zonations and scales. For each combination of zonation and scale, the authors estimated a set of parameters and then used regression to extrapolate the values of the parameters on a minimal unit area. They also constructed an interval for these parameters by simulating that according to the extrapolated values obtained in the first step. [9] examined the MAUP effect on UK census data, utilizing the predefined scales provided by the UK census. Their methodology involves calculating the difference in correlation between pairs of variables across different scales and subsequently testing the significance of these differences. It's worth noting that their approach does not account for variations in scales or zoning systems. [8] introduced a non-parametric hypothesis test known as S-MAUP to test if a dataset is suitable to the MAUP. The null hypothesis is that the dataset has no MAUP effects. The method empirically derives the distribution and critical values.

Chapter 2

Methodology

2.1 Overview of our method

2.1.1 Definitions and notation

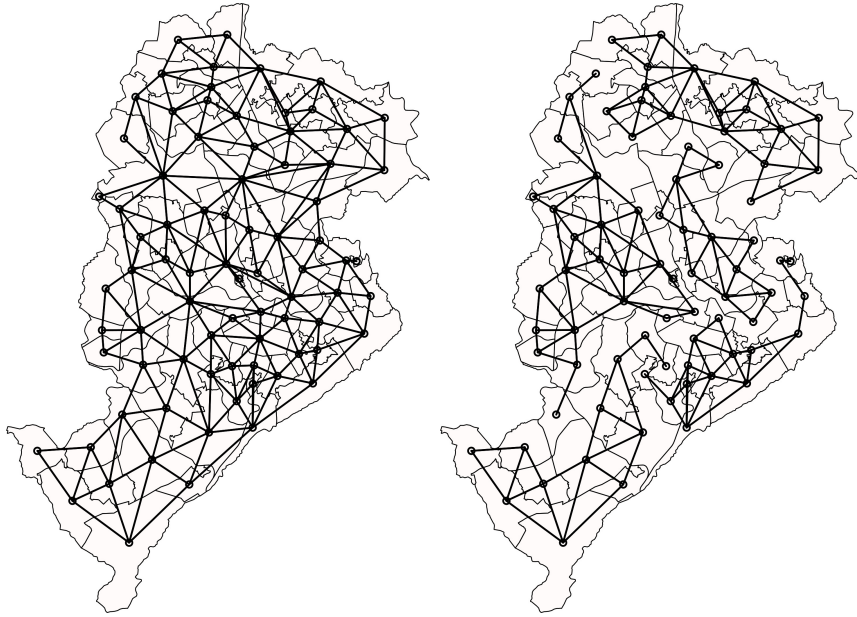
We represent a region partitioned into n small areas by an undirected graph $\mathcal{G} = (V, E)$ where the centroids of the areas are the n vertices or nodes in the set V and E is the set of edges. There is an edge connecting vertices v_i and v_j if areas i and j share geographical boundaries. The left-hand side plot in Figure 2.1 shows a map transformed into such a graph. A *path* from node v_1 to node v_k is a sequence of nodes v_1, v_2, \dots, v_k that are connected by edges $(v_1, v_2), \dots, (v_{k-1}, v_k)$. A graph is said to be *connected* if, for any pair of nodes v_i and v_j there is at least one path connecting them. We assume that the adjacency neighborhood graphs are always connected graphs.

We define a *spatial cluster* or *zone* as any connected subset of nodes. The right-hand side plot in Figure 2.1 shows the adjacency graph on the left partitioned into 4 spatial clusters or zones. The graph is partitioned into s spatial clusters $\mathcal{G}_1, \dots, \mathcal{G}_s$ if the clusters are disjoint, their union is \mathcal{G} , and each one of them is a connected subgraph. The number s of spatial clusters is referred to as the scale parameter. The specific partitioning $\mathcal{G}_1, \dots, \mathcal{G}_s$ is denoted as the zoning and represented by $z = (\mathcal{G}_1, \dots, \mathcal{G}_s)$.

In area i , we have a k -dimensional feature vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ik}) \in \mathbb{R}^k$. For instance, \mathbf{x} might contain k socio-economic characteristics of each area. Using these attributes, we compute a statistic of interest $T(\mathbf{x}_1, \dots, \mathbf{x}_n)$, such as an inequality index of the first feature $\{x_{i1}, i = 1, \dots, n\}$, the correlation between the first two attributes in \mathbf{x} , or the linear regression of x_{i1} on the other variables in \mathbf{x} . The statistic computed in the disaggregated map is denoted by $T(n, (1, 2, \dots, n))$, where the second argument $(1, 2, \dots, n)$ represents the partitioning when each individual area is one zone.

Given a spatial zoning $z = (\mathcal{G}_1, \dots, \mathcal{G}_s)$, there is a feature vector $\mathbf{x}_{\mathcal{G}_i} = (x_{\mathcal{G}_i,1}, \dots, x_{\mathcal{G}_i,k})$ associated with the i -th zone. For example, the j -th variable in $\mathbf{x}_{\mathcal{G}_i}$ can be the mean of

Figure 2.1: Left: Map of an urban area transformed into a graph $\mathcal{G} = (V, E)$. Right: A spatial partitioning $z = (\mathcal{G}_1, \dots, \mathcal{G}_4)$ of the graph with 4 zones or spatial clusters.



Source: Developed by the author.

the j -th attribute over the subset of areas comprising \mathcal{G}_i :

$$x_{\mathcal{G}_i, j} = \frac{1}{n_i} \sum_{k \in \mathcal{G}_i} x_{k, j}$$

where n_i being the number of areas in \mathcal{G}_i . We denote by $T(s, z)$ the value of the statistic calculated with the zoning $z = (\mathcal{G}_1, \dots, \mathcal{G}_s)$ and the attributes $\mathbf{x}_{\mathcal{G}_i}$. That is,

$$T(s, z) = T(s, (\mathcal{G}_1, \dots, \mathcal{G}_s)) = T(\mathbf{x}_{\mathcal{G}_1}, \dots, \mathbf{x}_{\mathcal{G}_s})$$

We denote by $\mathcal{L}(s)$ the probability distribution of $T(s, z)$ when the zoning $z = (\mathcal{G}_1, \dots, \mathcal{G}_s)$ is selected randomly among all possible partitionings of the map into s spatial clusters. The random selection is made uniformly, with all possible zonings at s scale level having the same probability of being selected.

We define two functions:

- $M(s) = \max_z \{T(s, z)\} = \max_z \{T(\mathbf{x}_{\mathcal{G}_1}, \dots, \mathbf{x}_{\mathcal{G}_s})\}$, the maximum the statistic T can reach when we scan all possible z zonings for each scale level s .
- $m(s) = \min_z \{T(s, z)\} = \min_z \{T(\mathbf{x}_{\mathcal{G}_1}, \dots, \mathbf{x}_{\mathcal{G}_s})\}$, similarly, the minimum value for each s .

These functions can be plotted as curves and Figure 1.2 illustrates two examples of the enveloping curves $M(s)$ and $m(s)$. Given a number s of regions, the $M(s)$ and $m(s)$ enveloping curves are defined by evaluating $T(s, z)$ for all possible spatial partitions of the map into s regions.

As the number of partitions is explosive even for a moderate number n of areas, it is unfeasible in practice to obtain the probability distribution $\mathcal{L}(s)$ and the exact values for $M(s)$ and $m(s)$. We need some heuristics to obtain approximations for the distribution and for $M(s)$ and $m(s)$. Our method is based on the SKATER algorithm proposed by [1] and summarized in Section 2.1.2. This is an algorithm to partition a map into disjoint regions formed by aggregation of connected areas. SKATER optimizes the creation of homogeneous regions and it is extremely fast compared to alternative spatial partitioning methods [2].

2.1.2 The SKATER algorithm

The SKATER objective is to aggregate adjacent small areas into larger regions that are homogeneous with respect to the vector of attributes \mathbf{x} . With these attributes, we calculate a pairwise dissimilarity or cost measure between areas i and j : $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$. The pairwise cost can be simply a Euclidean distance between the k -dimensional vectors \mathbf{x}_i and \mathbf{x}_j if their component variables have compatible scales. Given a spatial cluster \mathcal{G} , let $\mu(\mathcal{G})$ be the mean vector of its component areas. SKATER has the objective of finding the partition $P_{\mathcal{G}} = \{\mathcal{G}_1, \dots, \mathcal{G}_C\}$ that minimizes the function

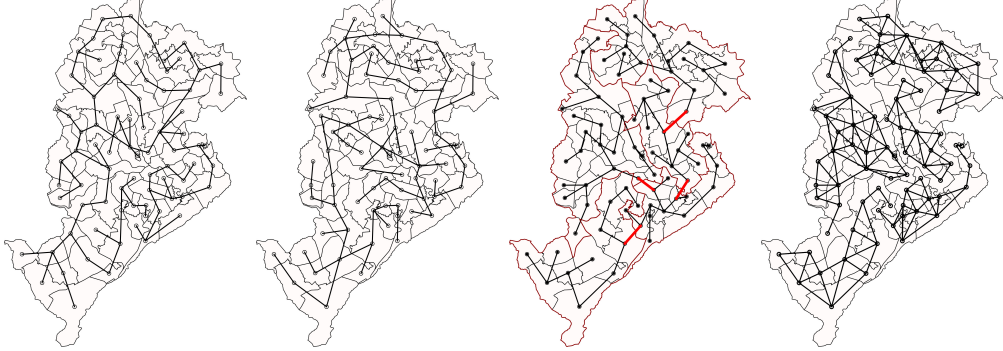
$$F(\mathcal{G}_1, \dots, \mathcal{G}_C) = \sum_{q=1}^C \sum_{i \in \mathcal{G}_q} d(\mathbf{x}_i, \mu(\mathcal{G}_q)). \quad (2.1)$$

It efficiently solves this problem by creating a minimum spanning tree (MST) $\mathcal{T} = (V, T)$ of the graph $\mathcal{G} = (V, E)$ based on the pairwise cost d_{ij} . This first step does not make use of the objective function in (2.1). This is made in the second step, when we successively and hierarchically prune the MST to obtain the spatial clusters minimizing F in (2.1).

Thanks to the spanning tree properties, the enumeration problem of this optimization is dramatically reduced. A spanning tree is a sub-graph of $\mathcal{G} = (V, E)$ containing all its n vertices in V and with only $n - 1$ edges from E . These edges are such that there is one, and only one, path between any two pairs of nodes of V . The left-hand side plot in Figure 2.2 shows two spanning trees associated with the adjacency graph $\mathcal{G} = (V, E)$ shown in Figure 2.1. These properties ensure that by pruning $c - 1$ edges from a spanning tree we immediately determine c spatial clusters in the original map. The rightmost plot in Figure 2.1 shows the spatial clusters determined by the pruning of the thick edges shown in the third plot.

If n is the number of nodes in V , then any spanning tree has $n - 1$ edges. Pruning

Figure 2.2: Left: Two spanning trees associated with the adjacency graph $\mathcal{G} = (V, E)$ shown in Figure 2.1. Third plot: Partitioning of the graph in 5 spatial clusters by pruning 4 edges from the spanning tree. The pruned edges are shown in red. Right: Spatial clusters determined by the pruning of the red edges in the previous spanning tree.



Source: Developed by the author.

the MST \mathcal{T} of k edges partition the graph into $k + 1$ spatial clusters. SKATER adopts a sequential procedure, selecting one edge at each iteration. Therefore, the problem becomes manageable as we need to sequentially select one edge to prune out of those still present in the MST.

For this pruning stage, we associate a different cost to each of the MST edges rather than the pairwise cost d_{ij} used to build the MST \mathcal{T} . Let $SSTO = \sum_i \|\mathbf{x}_i, \mu(\mathcal{G})\|$ be the dissimilarity between the areas \mathbf{x}_i and the average attribute profile in the graph \mathcal{G} , where $\|\cdot\|$ denotes the Euclidean norm. Let \mathcal{G}_1 and \mathcal{G}_2 be the spatial clusters resulting from the removal of one edge from \mathcal{T} . Consider

$$SSW = \sum_q \sum_{i \in \mathcal{G}_q} \|\mathbf{x}_i - \mu(\mathcal{G}_q)\| \quad q = 1, 2, \dots, Q. \quad (2.2)$$

The smaller the value of SSW , the more homogeneous the resulting spatial clusters. Define the cost of removing the edge as equal to $SSTO - SSW$ and hence an edge with a large cost indicates that its removal results in homogeneous spatial clusters.

The procedure is iterated until some stopping rule is met. For example, until any additional partitioning results in spatial clusters with population less than a critical threshold. Another possible stopping rules are: to stop if the decrease of the objective function (2.1) is small or if a desired number s of spatial clusters is reached. Summarizing the entire procedure, the SKATER algorithm is given by Algorithm 2.1.

The SKATER algorithm complexity is $\mathcal{O}(|V|^2 \log |V|)$. This is because the minimum spanning tree stage is executed in $\mathcal{O}(|E| + |V| \log |V|)$ running times, using Prim's algorithm with a Fibonacci Heap [7]. Stage 2 requires $\mathcal{O}(|V|^2 \log |V|)$ and therefore the total execution time is $\mathcal{O}(|E| + |V| \log |V|) + \mathcal{O}(|V|^2 \log |V|) = \mathcal{O}(|V|^2 \log |V|)$. A comparison between SKATER and other spatial partitioning methods was provided in [2], concluding that it yields excellent results in terms of creating homogeneous regions in a much shorter execution time compared to the alternatives.

Algorithm 2.1: SKATER

Data: graph $\mathcal{G} = (V, E)$ with attribute values $\{\mathbf{x}_v\}$ for $v \in V$
Result: Spatial partition $P_{\mathcal{G}} = \{\mathcal{G}_1, \dots, \mathcal{G}_s\}$
// Calculate pairwise dissimilarities
 $d_{ij} \leftarrow \|\mathbf{x}_i - \mathbf{x}_j\|$
// Build a minimum spanning tree (MST)
 $\mathcal{T} \leftarrow MST(\mathcal{G})$
// Initialize the partition
 $P_{\mathcal{G}} \leftarrow \emptyset$
// Prune \mathcal{T} of $(u, v)^$ s.t. (2.2) is minimized*
 $(u, v)^* = \arg \min SSW$
 $P_{\mathcal{G}} \leftarrow \{\mathcal{G}_1, \mathcal{G}_2\}$
while *stopping rule is not met* **do**
 // Repeat the pruning step splitting the spatial cluster among
 $\mathcal{G}_1, \dots, \mathcal{G}_q$ *that minimizes (2.2)*
end
return $P_{\mathcal{G}}$

Algorithm 2.2: Obtaining random partitions

Data: graph $\mathcal{G} = (V, E)$ with attribute values $\{\mathbf{x}_v\}$ for $v \in V$
Result: Spatial partition $P_{\mathcal{G}} = \{\mathcal{G}_1, \dots, \mathcal{G}_s\}$
// Generate pairwise dissimilarities at random
 $d_{ij} \leftarrow generateRandomNumber$
// Build a minimum spanning tree (MST)
 $\mathcal{T} \leftarrow MST(\mathcal{G})$
// Initialize the partition
 $P_{\mathcal{G}} \leftarrow \emptyset$
// Prune \mathcal{T} of $(u, v)^$ at random*
 $P_{\mathcal{G}} \leftarrow \{\mathcal{G}_1, \mathcal{G}_2\}$
for 2 *to* n **do**
 // Repeat the pruning step splitting the spatial cluster among
 $\mathcal{G}_1, \dots, \mathcal{G}_q$ *at random*
end
return $P_{\mathcal{G}}$

2.2 Estimating $\mathcal{L}(s)$ and the extreme bounds $M(s)$, and $m(s)$

As in SKATER, we start with the adjacency graph between the regions. We assign independent and identically distributed continuous random variables d_{ij} as edges' weights. Next, an MST is created. Sequentially deleting one edge at a time from the MST we create a series of partitionings at the $n - 1, n - 2, \dots, 2$ scale levels. At each step of the sequence, we select a random edge from the tree to be deleted creating an

additional spatial cluster. At each scale level s and using the specific random partitioning z produced by the deletions, we calculate the statistic $T(s, z)$. The sequence $\{T(n, z_n), T(n-1, z_{n-1}), \dots, T(2, z_2)\}$ is connected and shown as a random trajectory in Figure 1.2. By repeating this process independently B times we generate several trajectories. The B simulated values at a fixed scale s , $\{T(s, z_s^{(1)}), \dots, T(s, z_s^{(B)})\}$, is a random sample from the $\mathcal{L}(s)$ probability distribution and they can be used to estimate the $\mathcal{L}(s)$ parameters, including $M(s)$ and $m(s)$. This process is described in the Algorithm 2.2.

On one hand, this is an algorithm that works regardless of the statistic $T(s, z)$. On the other hand, obtaining the estimates for $M(s)$ and $m(s)$ may require a substantial number of simulations (B). Hence, we propose a second method that, although not providing an estimate for the distribution $\mathcal{L}(s)$, estimates the extreme bounds $M(s)$ and $m(s)$ more directly. However, this second method is more restrictive because it is a costly sampling procedure, even if based on the fast SKATER algorithm to obtain the z partitionings.

To obtain $M(s)$ and $m(s)$ we propose one additional modification in the usual SKATER second stage. Instead of partitioning the MST at random, we will prune the edge that maximizes (or minimizes) $T(s, z)$. For that, we need to assign new costs to the tree edges when we split the MST. Remind a fundamental property of the spanning tree: removing any $s-1$ edges from the tree creates s spatial clusters determining a $z = (\mathcal{G}_1, \dots, \mathcal{G}_s)$ partitioning with the current value

$$T(s, z) = T(s, (\mathcal{G}_1, \dots, \mathcal{G}_s)) = T(\mathbf{x}_{\mathcal{G}_1}, \dots, \mathbf{x}_{\mathcal{G}_s})$$

for the statistic of interest (see Figure 2.2). At this step, we assign a new cost to each of the $n-s$ remaining edges in the tree. If an edge is removed, we split a given spatial cluster into two new ones, giving rise to a new partitioning z^* , the scale changes to $s+1$ and the statistic of interest assumes the value $T(s+1, z^*)$. The cost measure assigned to each edge of the current tree is $1/T(s+1, z^*)$ and it has different values at different edges. To prune the current tree, we select that edge that minimizes $1/T(s+1, z^*)$. To obtain $m(s)$ we select that one that minimizes $T(s+1, z^*)$.

The first split of the MST may require a different approach depending on the T statistic. Suppose, for example, that T is the correlation coefficient. Pruning the MST of one of its (i, j) edges results in two subtrees, \mathcal{T}_1 and \mathcal{T}_2 , and the corresponding two spatial clusters that define the zoning $z^{(i,j)} = (\mathcal{G}_1^{(i,j)}, \mathcal{G}_2^{(i,j)})$ with scale level $s=2$. After deleting an (i, j) edge from \mathcal{T} , we can calculate the statistic of interest with the two spatial regions $z^{(i,j)}$ determined by the two \mathcal{T}_1 and \mathcal{T}_2 subtrees. But the correlation coefficient based on any two points is always $+1$, -1 , or 0 . Therefore, we change the criterion for the first split. In this case, instead of maximizing (or minimizing) the correlation coefficient based on two points, we maximize the slope of the line passing through the two points, so the cost will be $C_{ij} = \frac{y_{i2} - y_{j1}}{x_{i2} - x_{j1}}$. Algorithm 2.3 displays this method.

Algorithm 2.3: Estimating $M(s)$ when $T(s, z)$ is the correlation

```

Data: graph  $\mathcal{G} = (V, E)$  with attribute values  $\{\mathbf{x}_v\}$  for  $v \in V$ 
Result: Spatial partition  $P_{\mathcal{G}} = \{\mathcal{G}_1, \dots, \mathcal{G}_s\}$ 
// Calculate pairwise dissimilarities as the orthogonal projection
 $d_{ij} \leftarrow generateRandomNumber$ 
// Build a minimum spanning tree (MST)
 $\mathcal{T} \leftarrow MST(\mathcal{G})$ 
// Initialize the partition
 $P_{\mathcal{G}} \leftarrow \emptyset$ 
// Prune  $\mathcal{T}$  of  $(u, v)^*$  that maximizes (or minimizes) the slope
 $(u, v)^* = \arg \max slope$ 
 $P_{\mathcal{G}} \leftarrow \{\mathcal{G}_1, \mathcal{G}_2\}$ 
for 2 to  $n$  do
    | // Repeat the pruning step splitting the spatial cluster among
    |    $\mathcal{G}_1, \dots, \mathcal{G}_q$  maximizing (or minimizes) the correlation
end
return  $P_{\mathcal{G}}$ 

```

Even though this method produces better estimates of $M(s)$ and $m(s)$, it still relies on simulations of multiple trajectories, which can still have a high computational cost. We also developed a single-shot estimator to $M(s)$ and $m(s)$ when $T(s, z)$ is the correlation coefficient. That is, the method does not require several simulations of $T(s, z)$, instead, it estimates the values of $M(s)$ and $m(s)$ only once. However, this method crucially depends on the selection of an appropriate dissimilarity measure d_{ij} for the SKATER first stage. This appropriate dissimilarity measure varies with the statistic $T(s, z)$.

To explain how this single-shot estimator works, we will focus on a specific $T(s, z)$ statistic, the usual Pearson correlation index between two variables X and Y measured in each region. The index is invariant by location and scale so we can consider the case where the variables are standardized with mean zero and unit standard deviation. In this case, the linear regression line has a slope equal to 1 or -1, and the most extreme correlation is obtained when all the (x_i, y_i) are perfectly aligned along one of these lines.

Consider the curve $M(s)$, which is the maximum value for the correlation coefficient at the s scale. If we aggregate the areas i and j , their points (x_i, y_i) and (x_j, y_j) will be substituted by the average (x_m, y_m) where $x_m = (x_i + x_j)/2$ and $y_m = (y_i + y_j)/2$. This average may be weighted by, for example, population sizes. The d_{ij} pairwise dissimilarity measure for the SKATER first stage is given by the orthogonal projection distance between the (x_m, y_m) mean point and the line $y = x$ (if the global correlation is positive) or $y = -x$ (if it is negative). Considering the positive case, the distance (and dissimilarity) is given by $d_{ij} = |y_m - x_m|/\sqrt{2}$. As the denominator is the same for all pairs, it can be ignored. We build a minimum spanning tree based on these d_{ij} dissimilarity measures. To obtain $m(s)$, we take the distance to the line associated with $y = -x$ which is associated with the most extreme negative correlation of the standardized variables. This implies that

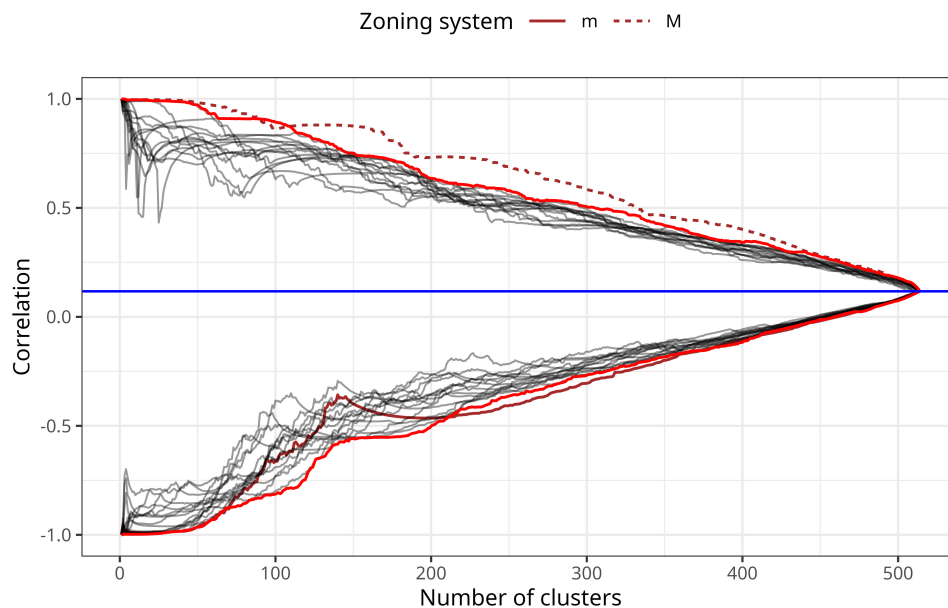
Algorithm 2.4: Estimating $M(s)$ in an efficient way when $T(s, z)$ is the correlation

```

Data: graph  $\mathcal{G} = (V, E)$  with attribute values  $\{\mathbf{x}_v\}$  for  $v \in V$ 
Result: Spatial partition  $P_{\mathcal{G}} = \{\mathcal{G}_1, \dots, \mathcal{G}_s\}$ 
// Calculate pairwise dissimilarities as the orthogonal projection
if  $\rho > 0$  then
  |  $d_{ij} \leftarrow |y_m - x_m|$ 
end
else
  |  $d_{ij} \leftarrow |y_m + x_m|$ 
end
// Build a minimum spanning tree (MST)
 $\mathcal{T} \leftarrow MST(\mathcal{G})$  // Initialize the partition
 $P_{\mathcal{G}} \leftarrow \emptyset$ 
// Prune  $\mathcal{T}$  of  $(u, v)^*$  that maximizes (or minimizes) the slope
 $(u, v)^* = \arg \max \text{slope}$   $P_{\mathcal{G}} \leftarrow \{\mathcal{G}_1, \mathcal{G}_2\}$  for 2 to  $n$  do
  | // Repeat the pruning step splitting the spatial cluster among
  |  $\mathcal{G}_1, \dots, \mathcal{G}_q$  maximizing (or minimizes) the correlation
end
return  $P_{\mathcal{G}}$ 

```

Figure 2.3: Comparing single-shot estimator (dotted line) with the estimates of M and m (solid red lines). Each black line represents an output of the algorithm 2.3 using the usual Pearson correlation coefficient as $T(s, z)$. Data is from Riverside County of 2020 ACS. The correlation was evaluated between the proportion of the black population and household income.



Source: Developed by the author.

$d_{ij} = |y_m + x_m|/\sqrt{2}$. This method is displayed at Algorithm 2.4. Figure 2.3 displays several outputs of Algorithm 2.3 and its efficient version (Algorithm 2.4). Note that the estimates produced by the optimized version are frequently close to the ones produced by the costly algorithm.

2.3 The MAUP index

Based on the distribution $\mathcal{L}(s)$ and the bounds $M(s)$ and $m(s)$, we introduce two different measures for the MAUP effect. The zoning effect is associated with the variation of the statistic T under different aggregations at the same scale level s . This is measured by the range $M(s) - m(s)$ which represents the interval within which we can create zonings at scale s . Consider the left-hand side plot in Figure 1.2. When $s = 100$, we can create zonings producing correlation varying from -0.81 to 0.77, approximately, while when $s = 200$, this variation has a smaller interval from -0.58 to 0.57. As this range fluctuates with s , the only way to obtain a single measure of the zoning effect is by summarizing the range $M(s) - m(s)$ over s . The most obvious way is by averaging the different values of $M(s) - m(s)$. By linearly interpolating between the discrete values of s , we can propose an index based on the area between the $M(s)$ and $m(s)$ curves. The larger this area, the larger the zoning effect. Considering the plots in Figure 1.2, we have a much larger zoning effect in the left-hand side plot than in the right-hand side plot. To obtain a reference value, we consider the maximum value this area could reach. This is given by the rectangle determined by horizontal lines at the extremes of $M(s)$ and $m(s)$. That is, with height given by $h = \max_s M(s) - \min_s m(s)$ and basis equal to the scale level range $n - 0$. Then, by interpolating the discrete-based $M(s)$ and $m(s)$, we define the MAUP effect I_z by integrating the zoning effect at scale s over all possible scales with an optional weighting function $w(s)$:

$$I_z = \frac{\int_0^n (M(s) - m(s))w(s)ds}{h \int_0^n w(s)ds} \quad (2.3)$$

The $w(s)$ weight function is useful if we want for discarding or downplaying unreasonable scales. Near the lower end of the scale range, when we have maps with a very small number of regions, such as $s = 2$ or 3 , we have the largest variation in the statistics. As shown in Figure 1.2, the theoretically extreme values $+1$ and -1 for the correlation coefficient are observed only very close to the lower end of the scale range. Similarly, s values very close to their maximum are likely to be of little interest. The reason is that, starting at the maximum scale $s = n$, the next smaller scale value, equal to $n - 1$, can

be obtained only by aggregating two adjacent areas leaving all the others as they were initially. This implies that the change in the T statistic is likely to be tiny, as can be observed again in the right-most extreme scale values in Figure 1.2. Maps formed by such a small number of regions or with practically the same regions as the original map are likely to be of little value and we may want to downweight these s values, concentrating the calculation of (2.3) on intervals of practical value. This is obtained with the weight function $w(s)$. For example, we may define $W(s) = 1$ only when $s \in (\delta, n - \delta^*)$ for some positive δ and δ^* .

As previously mentioned, the metric I_z encapsulates a pessimistic perspective, reflecting an analyst's determination to ensure the robustness of the analysis against arbitrary partitionings. This approach adopts an adversarial mindset, considering at each scale s the partitionings that could potentially exert the greatest influence on the statistic under examination. These partitionings represent the choices of an adversary seeking to disrupt the analysis based on the refined data, aiming to ascertain the maximum impact a partitioning could impose on the analysis at scale s .

We introduce a second metric to assess the MAUP effect, which takes a less pessimistic stance. A justification for this is that the pessimistic approach may be extreme. Figure 1.2 shows that the $M(s)$ and $m(s)$ curves represent extreme bounds with the large number of simulated trajectories staying far away from these limits. This implies that I_z is not measuring the typical effect of a chance partitioning of the space but rather a purposely partitioning aiming at driving the statistic to its maximum or minimum values. In this sense, these extreme partitionings are not representative of what may occur if the new geography is not intentionally planned to maximally change T . The calculation of this metric is similar to the previous, but instead of focusing on the extreme zoning systems obtained using the Algorithm 2.4, we use as $M(s)$ and $m(s)$ in the Equation 2.3 the maximum and minimum of $T(s, z)$ observed in the random trajectories (the red lines in the Figure 1.2). We denote this metric as I_m .

2.3.1 The scale effect

The scale effect is related to the variation of the statistic T due to a change in the scale that the data was aggregated, keeping the same zoning system. To access the scale effect, instead of focusing on the maximum potential impact of partitionings at any scale s , one could consider the average maximum impact achievable at each scale s . A random trajectory is selected, and its value $T(s, z)$ at scale level s is examined under a specific partitioning z , then the maximum difference between $T(s, z)$ and the extreme bounds at

the next scale level would measure the change in the statistic due to a change in a single scale for a fixed zoning system z : $\max(|M(s-1) - T(s, z)|, |m(s-1) - T(s, z)|)$. This represents how much the statistic T can vary due to a single change in the scale level. The two differences are the slopes of the line that connects an extreme bound at $s-1$ and $T(s, z)$. So, the scale effect is measured by the average of these differences over the zoning systems:

$$I_s = \frac{1}{Z \times (n-1)} \sum_{z=1}^Z \sum_{s=2}^n \max(|M(s-1) - T(s, z)|, |m(s-1) - T(s, z)|), \quad (2.4)$$

where Z is the total of different zoning systems. There is no upper limit for this metric.

One important point is that the scale and zoning effects are tangled. When we average over different scale levels s , or over different zoning systems z , we mix up the zoning effect with the scale effects. We will show in the section 3.1.3 that the correlation between I_s and I_z is close to 1. This way, these two measures carry the same information and should use only one of them.

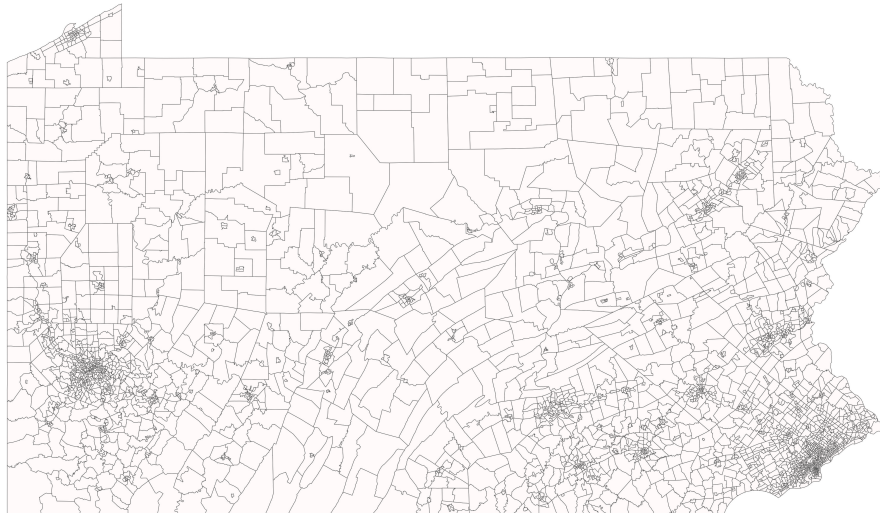
Chapter 3

Results

3.1 Empirical analysis

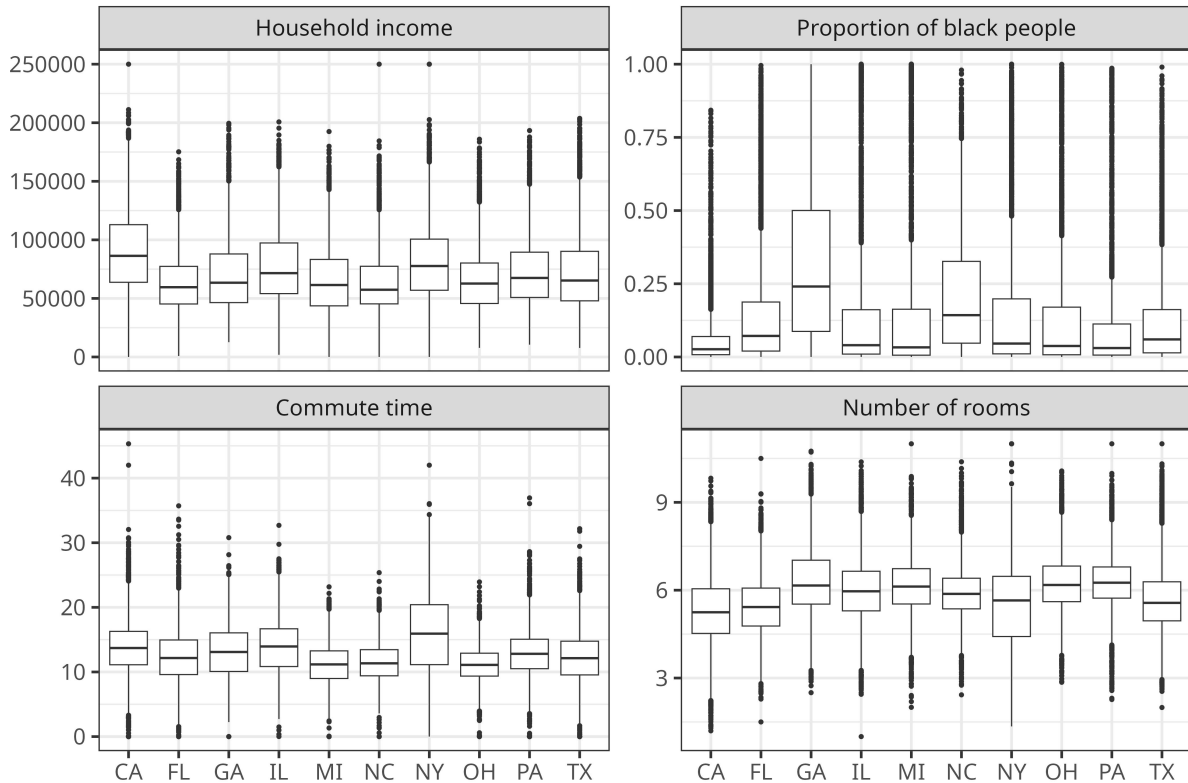
In this section, we show the performance of our measure in a large number of datasets. We start with a brief description of the data we used and how it is collected, and then we proceed to the results.

Figure 3.1: Map of Pennsylvania Divided into ACS Census Tracts



Source: Developed by the author.

Figure 3.2: Boxplot of each variable across the states.



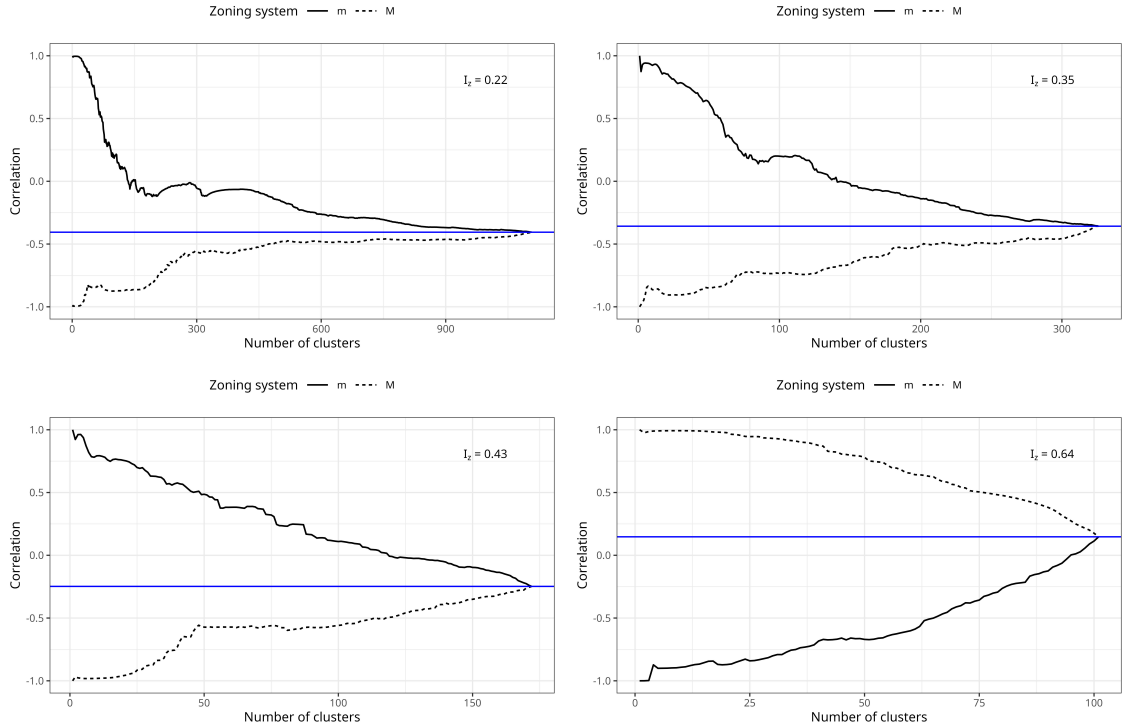
Source: Developed by the author.

3.1.1 Datasets

We utilized data from the 2020 American Community Survey (ACS) for our analyses, measured at the Census tract level, which was subsequently aggregated up to the county level. Our dataset encompassed counties in California (CA), Florida (FL), Georgia (GA), Illinois (IL), Michigan (MI), New York (NY), North Carolina (NC), Ohio (OH), Pennsylvania (PA), and Texas (TX). Figure 3.1 displays the state of Pennsylvania divided into census tracts. Counties with fewer than 20 census tracts were excluded, resulting in a total of 319 counties. The number of Census tracts varied from 23 to 2,460 (in Los Angeles), with the first quartile, median, and third quartile equal to 32, 52, and 115, respectively. We used the following variables: average commute time (in minutes), average household income (in American dollars), percentage of the Black population, and average number of rooms in the residence.

The data was collected using the `tidycensus` [31] R [25] package. Some of the variables were provided as counts within ranges of values. For instance, the annual income is given in several ranges (0 - 5,000\$, 5,000\$ - 10,000\$, *etc*). In these cases, we calculated the mean value of each range and multiplied it by the number of observations in that

Figure 3.3: Typical behaviour of $M(s)$ and $m(s)$. The solid blue line indicates the correlation at the original scale. Upper left: Harris County (TX) using household income and proportion of black population. Upper right: Franklin County (OH) using proportion of the black population and number of rooms. Lower left: San Joaquin County (CA) using household income and proportion of black population. Lower right: Washtenaw County (MI) using commuting time and household income.

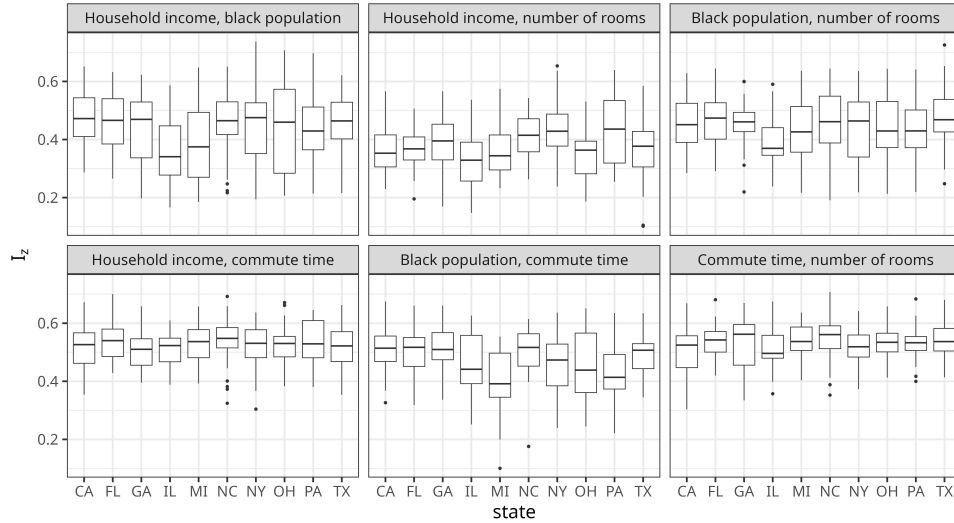


Source: Developed by the author.

range and tract. We then divided by the total population of the tract to approximate the mean value. Figure 3.2 shows boxplots for each variable at the census tract level for each state analyzed.

3.1.2 Results

To validate the methodology proposed in the previous sections, we analyzed the correlation coefficient between the specified pairs of variables. In the pessimistic scenario (Algorithm 2.4), our methodology was applied to each county and each pair of variables, leading to a total of 1914 distinct scenarios (319 counties with 6 pairs of variables each). Due to computational constraints, we used only 30 counties for each pair of variables for the random scheme (Algorithm 2.2). For each scale, we simulated 100 random zoning systems. The results are available on an interactive platform de-

Figure 3.4: Boxplot of I_z for each pair of variables and state.

Source: Developed by the author.

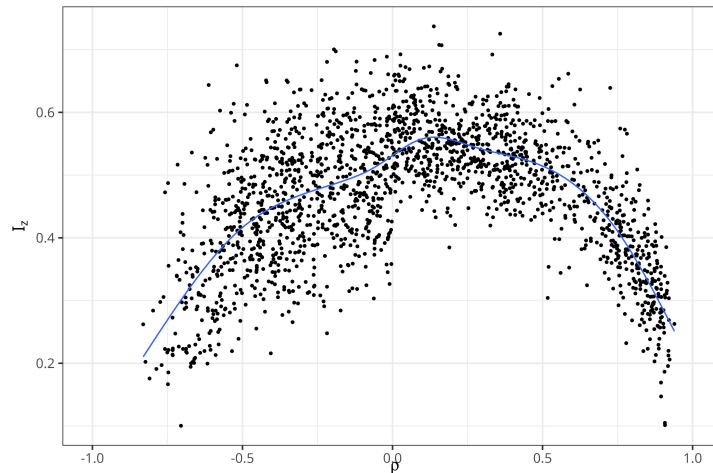
veloped by the authors using the Shiny R package [5], accessible via the following link: <https://vriffel.shinyapps.io/maup/>. Figure 3.3 displays the typical behaviours of $M(s)$ and $m(s)$. Note that as the scale of aggregation approaches the original data, the extreme boundaries align more closely with the original value of $T(n)$. When $|T(n)|$ is large, one of the boundaries is constrained by its value, thereby reducing the MAUP effect.

The boxplot in Figure 3.4 illustrates the distribution of I_z for each pair of variables across different states. Overall, I_z varied around 0.4 and 0.55. The analysis reveals that the correlation between average commute time to work and household income experienced more pronounced MAUP effects, whereas the correlation between household income and the number of rooms witnessed fewer MAUP influences. There is noticeable variation in the distribution of I_z for different pairs of variables across states. For instance, the median I_z for the correlation between the black population and the number of rooms in Illinois was 0.37, whereas for other states, this value was approximately 0.45.

We also plotted I_z for each pair of variables in each county against their respective correlation coefficients, as seen in Figure 3.5. The blue line shows a smooth of the I_z values adjusted by the LOESS method. The data indicate that the MAUP effect correlates with the value of the statistics obtained at the original scale. Statistics closer to -1 or 1 exhibit a less pronounced MAUP effect, while those nearer to 0 show a greater MAUP effect. Also, when the correlation at the original scale was closer to zero, the variance of I_z seems to be greater.

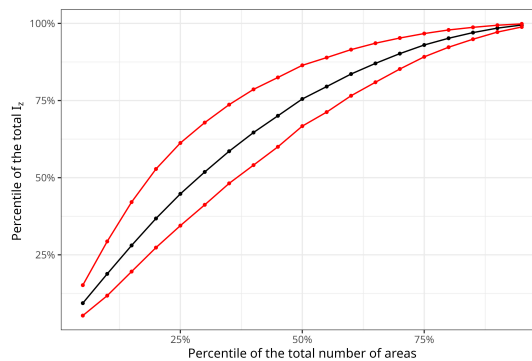
When we analyze the typical behaviour of $M(s)$ and $m(s)$ (Figure 3.3), we see that the effect of MAUP is more concentrated at smaller scale values. That is, we observe more extreme values for $M(s)$ and $m(s)$ when the scale is lower. To investigate this, Figure 3.6 displays the cumulative effect of I_z on the respective percentiles at the original scale.

Figure 3.5: Plot of I_z for each county and pair of variables against its respective correlation at the original scale. The blue line is the estimates using the LOESS method.



Source: Developed by the author.

Figure 3.6: Cumulative effect of I_z on the respective percentiles at the original scale. The black line is the percentile average of accumulated I_z . The red lines indicate the 5th and 95th of the accumulated I_z .



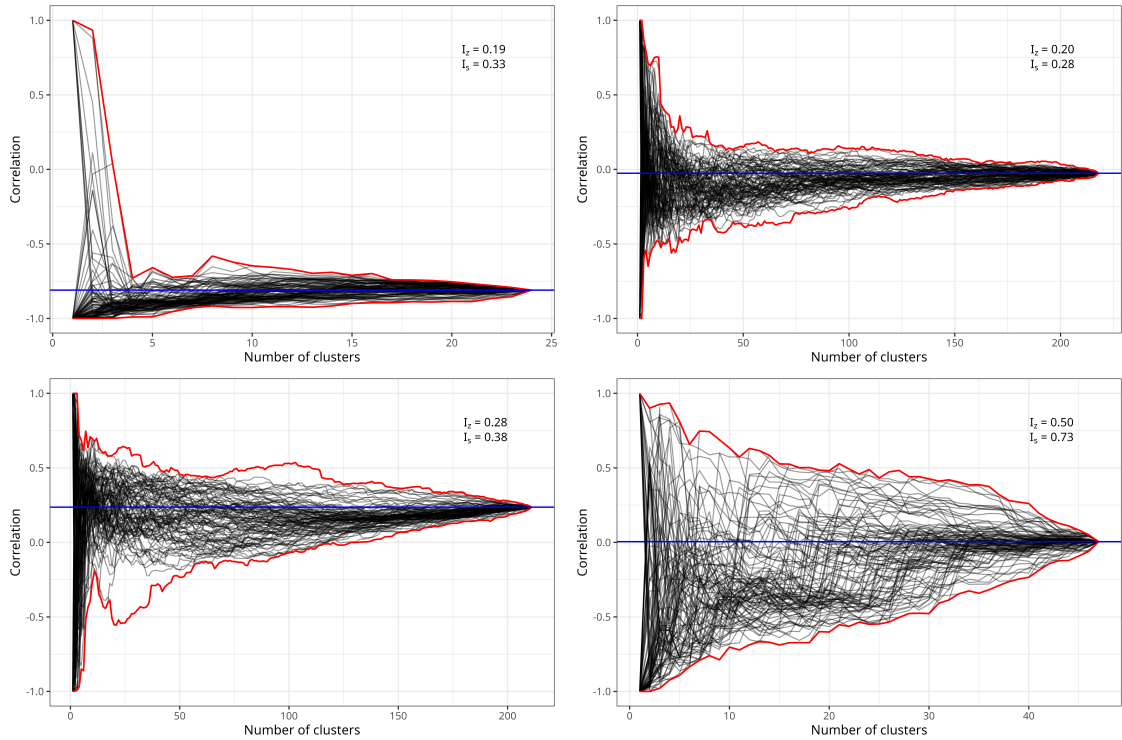
Source: Developed by the author.

Here, the black line represents the mean accumulated I_z , while the red lines denote the 5th and 95th percentiles. In the median, we see that almost 50% of the MAUP was found in around 25% of the total number of areas.

3.1.3 Random scenario

As previously, we display in Figure 3.7 four typical scenarios generated by different zoning systems. These plots reveal that aggregations can be more volatile in certain scenarios than in others, indicated by a larger variance in $\mathcal{L}(s)$. For instance, comparing the two plots on the right shows that the lower figure exhibits a significantly greater

Figure 3.7: Typical behaviour on the random scenario. Each black line represents a simulation from $\mathcal{L}(s)$. The solid blue line indicates the correlation at the original scale. The solid red line indicates the maximum and minimum at each scale. Upper left: Wilson County (NC) using commuting time and proportion of black population. Upper right: DuPage County (IL). Lower left: Lee County (FL). Lower right: Union County (NC), the latter three the correlation was calculated between commuting time and household income.



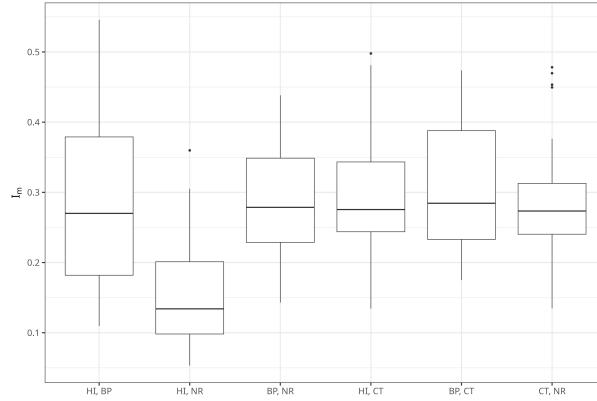
Source: Developed by the author.

variation in the values of $T(s, z)$, a characteristic captured by I_s .

The boxplot in Figure 3.8 shows the distribution of I_m values obtained under the random scheme. Once again, we observe numerous high values for the index. This raises concerns that different indexes might lead to varying conclusions when applied to the same data. As anticipated, the values of I_z (the pessimistic scenario) tend to be higher than those in the random scheme (I_m). Figure 3.9 compares the MAUP effect values obtained in both the pessimistic and random schemes. With a correlation coefficient of 0.64 between the two measurements, it indicates that both methods generally deliver consistent conclusions. Additionally, it is noteworthy that in only a few instances did the pessimistic scheme yield lower values than the random scheme.

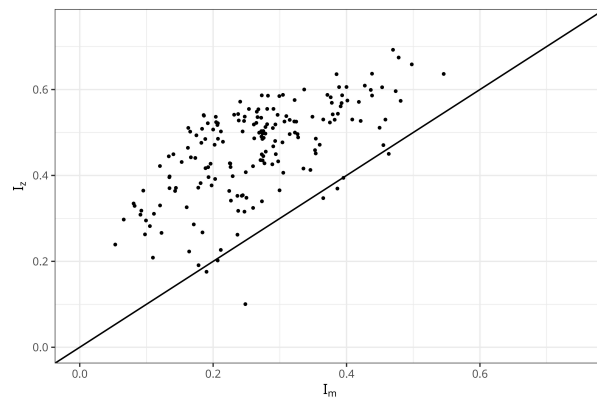
Finally, Figure 3.10 illustrates the relationship between I_z and I_s . This plot clearly indicates a high correlation between the scale effect and the zoning effect, suggesting that using both I_z and I_s may be redundant for measuring MAUP in a dataset. The correlation between I_z and I_s was found to be 0.99.

Figure 3.8: Boxplot of I_m for each pair of variables. HI stands for household income, BP for the proportion of the Black population, NR for the number of rooms and Ct for commuting time.



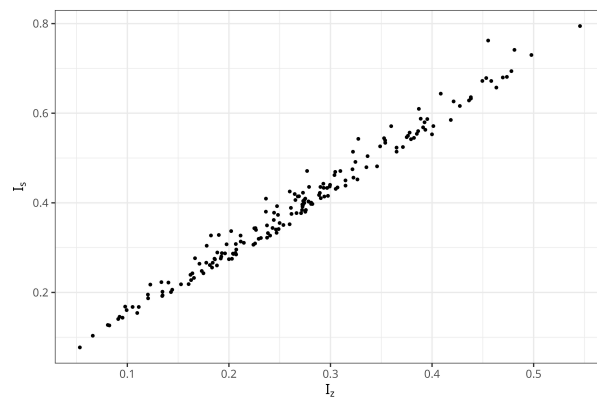
Source: Developed by the author.

Figure 3.9: Plot of both MAUP measurements. The solid line denotes the identity line.



Source: Developed by the author.

Figure 3.10: Plot of the scale effect (I_s) and the zoning effect (I_z). Both of measurements are highly correlated.



Source: Developed by the author.

Chapter 4

Conclusions

In this study, we introduce a novel stochastic approach to quantify the effects of the Modifiable Areal Unit Problem (MAUP) in spatial analyses. Our findings indicate that while MAUP can significantly impact certain analyses, its influence is not universal and varies with the context and data granularity. We propose an index that is both easy to use and interpret in practice. Although we utilized the correlation coefficient to derive our results, the proposed method is versatile and applicable to any spatial analysis.

The indices we developed offer a robust metric for evaluating the sensitivity of analyses to MAUP. Our observations reveal that the most significant variations due to MAUP occur at smaller scales, where the aggregation of areas has a more pronounced effect on the calculated statistics. Additionally, we found that scale and zoning effects seem to be intertwined.

While our methods provide a solid foundation for evaluating the MAUP, they also have limitations. For instance, the need for significant computational power to calculate the extremes $M(s)$ and $m(s)$ in large datasets can be a challenge. Future work may focus on algorithmic optimizations or the development of more efficient methods for estimating these limits.

Additionally, exploring the application of our metrics in different statistics and contexts, such as environmental studies and public health, could provide further insights into the applicability and robustness of our methods. Additional work is also needed to thoroughly investigate the index and establish a threshold at which MAUP becomes problematic.

References

- [1] Renato M Assunção, Marcos Corrêa Neves, Gilberto Câmara, and Corina da Costa Freitas. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811, 2006.
- [2] Orhun Aydin, Mark V Janikas, Renato Martins Assunção, and Ting-Hwan Lee. A quantitative comparison of regionalization methods. *International Journal of Geographical Information Science*, 35(11):2287–2315, 2021.
- [3] Thomas Blaschke, Geoffrey J Hay, Maggi Kelly, Stefan Lang, Peter Hofmann, Elisabeth Addink, Raul Queiroz Feitosa, Freek Van der Meer, Harald Van der Werff, Frieke Van Coillie, et al. Geographic object-based image analysis—towards a new paradigm. *ISPRS journal of photogrammetry and remote sensing*, 87:180–191, 2014.
- [4] Álvaro Briz-Redón. A bayesian shared-effects modeling framework to quantify the modifiable areal unit problem. *Spatial Statistics*, 51:100689, 2022.
- [5] Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. *shiny: Web Application Framework for R*, 2024. R package version 1.8.1.1.
- [6] Gang Chen, Qihao Weng, Geoffrey J Hay, and Yinan He. Geographic object-based image analysis (geobia): Emerging trends and future opportunities. *GIScience & Remote Sensing*, 55(2):159–182, 2018.
- [7] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2022.
- [8] Juan C Duque, Henry Laniado, and Adriano Polo. S-maup: Statistical test to measure the sensitivity to the modifiable areal unit problem. *PloS one*, 13(11):e0207377, 2018.
- [9] Robin Flowerdew. How serious is the modifiable areal unit problem for analysis of english census data? *Population trends*, 145:106–118, 2011.
- [10] A Stewart Fotheringham and David WS Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23(7):1025–1044, 1991.

-
- [11] Feng Gao, Shaoying Li, Zhangzhi Tan, Zhifeng Wu, Xiaoming Zhang, Guanping Huang, and Ziwei Huang. Understanding the modifiable areal unit problem in dockless bike sharing usage and exploring the interactive effects of built environment factors. *International Journal of Geographical Information Science*, 35(9):1905–1925, 2021.
- [12] Charles E Gehlke and Katherine Biehl. Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29(185A):169–170, 1934.
- [13] Andrew Gelman and Gary King. Estimating the electoral consequences of legislative redistricting. *Journal of the American statistical Association*, 85(410):274–282, 1990.
- [14] Andrew Gelman and Gary King. A unified method of evaluating electoral systems and redistricting plans. *American Journal of Political Science*, pages 514–554, 1994.
- [15] Reyhane Javanmard, Jinhyung Lee, Junghwan Kim, Luyu Liu, and Ehab Diab. The impacts of the modifiable areal unit problem (maup) on social equity analysis of public transit reliability. *Journal of Transport Geography*, 106:103500, 2023.
- [16] Maja Kucharczyk, Geoffrey J Hay, Salar Ghaffarian, and Chris H Hugenholtz. Geographic object-based image analysis: a primer and future directions. *Remote Sensing*, 12(12):2012, 2020.
- [17] Andrew B Lawson. *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. Chapman and Hall/CRC, 2018.
- [18] Andrew B Lawson, Fiona LR Williams, and Fiona Williams. *An introductory guide to disease mapping*, volume 10. John Wiley New York, 2001.
- [19] Duncan Lee, Chris Robertson, Colin Ramsay, and Kate Pyper. Quantifying the impact of the modifiable areal unit problem when estimating the health effects of air pollution. *Environmetrics*, 31(8):e2643, 2020.
- [20] David Manley. Scale, aggregation, and the modifiable areal unit problem. In *Handbook of regional science*, pages 1711–1725. Springer, 2021.
- [21] Duane F Marble. Some thoughts on the integration of spatial analysis and geographic information systems. *Journal of Geographical Systems*, 2:31–35, 2000.
- [22] S Openshaw and P Taylor. A million or so correlated coefficients. *Statistical Applications in the Spatial Sciences*, N. Wrigley (ed.). London, UK: Pion, 127:144, 1979.
- [23] Stan Openshaw. The modifiable areal unit problem. *Concepts and techniques in modern geography*, 1984.

-
- [24] S H Putman and S-H Chung. Effects of spatial system design on spatial interaction models. 1: The spatial system definition problem. *Environment and Planning A: Economy and Space*, 21(1):27–46, 1989.
- [25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.
- [26] Mark E Rush. *Does redistricting make a difference?: partisan representation and electoral behavior*. Lexington Books, 2000.
- [27] Matthew Tuson. *New geographical and statistical methods to address the modifiable areal unit problem, with applications in health*. PhD thesis, The University of Western Australia, 2022.
- [28] Matthew Tuson, Matthew Yap, Mei Ruu Kok, Bryan Boruff, Kevin Murray, Alistair Vickery, Berwin A Turlach, and David Whyatt. Overcoming inefficiencies arising due to the impact of the modifiable areal unit problem on single-aggregation disease maps. *International journal of health geographics*, 19:1–18, 2020.
- [29] Matthew Tuson, Matthew Yap, Mei Ruu Kok, Kevin Murray, B Turlach, and David Whyatt. Incorporating geography into a new generalized theoretical and statistical framework addressing the modifiable areal unit problem. *International journal of health geographics*, 18:1–15, 2019.
- [30] Matthew Tuson, Matthew Yap, and David Whyatt. Investigating local variation in disease rates within high-rate regions identified using smoothing. *Geospatial Health*, 18(1), 2023.
- [31] Kyle Walker and Matt Herman. *tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames*, 2024. R package version 1.6.3.
- [32] Lance A Waller and Bradley P Carlin. Disease mapping. *Chapman & Hall/CRC handbooks of modern statistical methods*, 2010:217, 2010.
- [33] Pengpeng Xu, Helai Huang, and Ni Dong. The modifiable areal unit problem in traffic safety: Basic issue, potential solutions and future research. *Journal of traffic and transportation engineering (English edition)*, 5(1):73–82, 2018.
- [34] Xiang Ye and Peter Rogerson. The impacts of the modifiable areal unit problem (maup) on omission error. *Geographical Analysis*, 54(1):32–57, 2022.
- [35] G. U. Yule and M. G Kendall. *An introduction to the theory of statistics*. Charles Griffin & Co. Ltd., London, 1950.

-
- [36] Wei Zeng, Chengqiao Lin, Juncong Lin, Jincheng Jiang, Jiazhi Xia, Cagatay Turkey, and Wei Chen. Revisiting the modifiable areal unit problem in deep traffic prediction with visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):839–848, 2021.