Efficient regionalization for spatially explicit neighborhood delineation

Ran Wei, Sergio J. Rey, and Elijah Knaap

Abstract

Neighborhood delineation is increasingly relied upon in urban social science research to identify the most appropriate spatial unit. However, existing approaches for neighborhood delineation are either nonspatial or lead to noncontiguous or overlapping regions. In this paper, we propose the use of max-p-regions for neighborhood delineation so that the geographic space can be partitioned into a set of homogeneous and geographically contiguous neighborhoods. In addition, we developed a new efficient algorithm to address the computational challenges associated with solving the max-p-regions so that it can be applied for large-scale neighborhood delineation. This new algorithm is implemented in the open-source Python Spatial Analysis Library (PySAL). Computational experiments based on both simulated and realistic data sets are performed and the results demonstrate its effectiveness and efficiency.

Introduction

A fundamental issue in the operationalization of many urban social science investigations is the choice of the spatial unit that organizes the socioeconomic data. In research on U.S. urban areas, the census tract has been taken as the common denominator as it represents a trade-off in favor of a larger number of socioeconomic variables at the cost of coarser spatial resolution relative to the increased spatial resolution of block-groups or blocks which come with a concurrent drop in the number of socioeconomic variables reported. In the classic approach to geodemographic analysis (e.g., Harris, et al. v2005) of urban neighborhoods, multivariate clustering algorithms are applied to census data to form neighborhood cluster types. The resulting neighborhood types are mapped, however, there is no guarantee that neighborhood types are spatially compact or contiguous. As a result, neighborhood types can be spatially fragmented which runs counter to the substantive understanding of neighborhoods as organizational units for human spatial behavior.

Despite the widespread adoption of tracts as the unit of choice, a number of recent calls have been made for more flexible approaches to the definition of neighborhoods supporting urban social science research. Clark et al. (2015) have argued persuasively for the notion of “bespoke” neighborhoods that can be formed by considering increasing distance buffers around an individual primitive unit (i.e., block or block group). The idea is to allow neighborhoods to be defined relative to a focal unit, and thus allow for the detection of scale effects reflected in different ethnic groups having neighborhoods of different extents. Spielman and Singleton (2015) also argued that the identification of bespoke neighborhoods by multivariate clustering can be one solution to the problem of large margins of error in the America Community Survey (ACS) data.

Parallel to the work on bespoke neighborhoods, there have been a number of recent advances in the longitudinal analysis of neighborhood dynamics in the geodemographic literature. An exemplar of this work is seen in Delmelle (2016) where four sets of decadal U.S. Census data (1980, 1990, 2000, 2010) are pooled to support the application of a multivariate clustering algorithm to define neighborhood typologies. Once the neighborhood types are identified, the resulting clusters from each time period are

mapped to delineate the neighborhood that each tract is assigned to. The dynamics come into play through a Markov chain based summarization of the transitions of individual tracts across neighborhood types over time. The sequence of transitions that each census tract experiences are then used in a second clustering exercise to identify a typology of neighborhood change.

The bespoke and longitudinal dynamics approaches towards neighborhood delineation offer a richer lens to study the spatial structure of urban areas over ones that treat census tracts and neighborhoods as identical. However, longitudinal approaches to neighborhood delineation are nonspatial in the sense that the only geographical information about a tract that is employed in the clustering is limited to its home metropolitan area. Like the geodemographic approach, the resulting neighborhoods are not guaranteed to be spatially contiguous. Its relative or absolute location within the urban context plays no formal role in the clustering algorithm. Given that many social attributes are spatially autocorrelated, that is, the local attributes affect the occurrence of the same phenomenon in neighboring areas, and local multicollinearity arises in many social phenomena, that is, different attributes are interdependent with each other (Openshaw and Taylor 1979; Getis and Ord 1992; Anselin 1995; Garreton and Sanchez 2016), such spatial information should play an explicit role in the clustering process and be integrated into the neighborhood delineation process. While the bespoke approach to neighborhood delineation does take the absolute and relative location of a focal tract into account, it does not result in a mutually exclusive partition of the tracts, as the resulting neighborhoods overlap.

There have been some clustering algorithms that explicitly account for intra-metropolitan spatial information and provide for exhaustive and mutually exclusive definition of neighborhoods. These clustering algorithms, generally categorized as regionalization methods, aim to partition the geographic space into a set of homogeneous and geographically contiguous regions (Openshaw and Rao 1995; Duque et al. 2007; Guo and Wang 2011; Garreton and Sanchez 2016). While the definition of neighborhood varies across disciplines, it typically refers to "a contiguous territory defined by a bundle of social attributes that distinguish it from surrounding areas" (Spielman and Logan 2013), coinciding with the goal of regionalization approaches (Folch and Spielman 2014). However, one of the major difficulties in applying regionalization methods to neighborhood delineation is their significant computational complexity (Spielman and Logan 2013). In this paper, we focused on one of the most widely used regionalization method, max-p-regions (Duque et al. 2012), and proposed a new efficient algorithm to address the computational challenges associated with solving it. In the next section, we provide a review of existing regionalization approaches with a particular focus on max-p-regions. Next, the new solution algorithm is presented. Finally, the proposed approach is applied to identifying neighborhood in several simulated datasets and census datasets, highlighting the effectiveness and efficiency of the new regionalization approach.

Regionalization

The need to aggregate spatial units into a set of contiguous regions arise in many social and environmental contexts, such as political districting, school districting, police patrol districting, habitat delineation, and various zone aggregation for modeling purpose. Many regionalization algorithms have been developed to fulfill such needs. For instance, Duque et al. (2011) formulated a typical regionalization problem as a mixed integer programming (MIP) model that can be solved using general MIP solver, like GUROBI (Gurobi 2019) or GLPK (GNU 2012). Guo (2008) integrated contiguity constraints into hierarchical clustering and developed the regionalization algorithm with dynamically constrained agglomerative clustering and partitioning (REDCAP). Li et al. (2014) developed a heuristic method,

memory-based randomized greedy and edge reassignment (MERGE), to aggregate spatial units into p compact and contiguous regions. A detailed review on regionalization algorithms can be found in Duque et al. (2007) and Garreton and Sanchez (2016).

Most of these regionalization algorithms require a prespecification of the number of regions identified (Folch and Spielman 2014; Garreton and Sanchez 2016). For example, the number of identified regions, *p*, is an input parameter for the p-regions model formulated in Duque et al. (2011), p-functional-regions formulated in Kim et al. (2013), and p-compact-regions formulated in Li et al. (2014). The users must select the level to cut for the hierarchical clustering based method like REDCAP in Guo (2008) and Guo and Wang (2011). However, the users rarely know the number of regions *a priori*. Alternatively, the max-p-regions proposed in Duque et al. (2012) allows the users to specify criteria that define a region and a regionalization scheme that satisfies the criteria is identified by solving the model. Such endogenization of the number of regions based on user-specified criteria make the max-p-regions approach ideally suited to identifying neighborhoods for further statistical modeling purpose (Folch and Spielman 2014). Here we reviewed max-p-regions model to highlight this and provide basis for the solution algorithm developed. Consider the following notation (Duque et al. 2012):

Parameters

$$i, j = index\ of\ spatial\ units, i \in I$$

$$k = index\ of\ potential\ regions, k \in K$$

$$c = index\ of\ contiguity\ order$$

$$d_{ij} = dissimilarity\ relationships\ between\ units\ i\ and\ j$$

$$l_i = spatially\ extensive\ attribute\ value\ of\ unit\ i$$

$$T = minimum\ value\ for\ attribute\ l\ at\ regional\ scale$$

$$w_{ij} = \{1, if\ unit\ i\ and\ j\ share\ a\ border\ 0, otherwise$$

$$N_i = \{w_{ij} = 1\}$$

$$F = 1 + \lfloor log\ log\ (\sum_i \quad \sum_j \quad d_{ij}) \rfloor$$

Decision variables:

$$y_{ij} = \{1, if\ units\ i\ and\ j\ belong\ to\ the\ same\ region\ 0, otherwise$$

$$x_i^{kc} = \{1, if\ unit\ i\ is\ assigned\ to\ region\ k\ in\ order\ c\ \ 0, otherwise$$

As the number of identified regions is unknown, the potential regions are indexed by k, which could range from 1 to the total number of spatial units. The contiguity order, indexed by c, is used to ensure contiguity within one region. Specifically, each region has only one root unit with a contiguity order $c = 0$. The other units that are assigned to the same region are either adjacent to the root unit, or next to a unit that has joined the region with a smaller order number. In addition to the attributes that are used to describe dissimilarity between units, the spatially extensive attribute, l_i, defines the size criteria that

each region must satisfy, such as the number of population and number of housing units. The number of regions becomes endogenous by ensuring each region exceed the threshold, T, on attribute l. The w_{ij} defines whether units i and j are adjacent, and the N_i is the set of units that are adjacent to unit i. Given this notation, the max-p-regions can be formulated as follows:

$$\left(-\sum_{k}\sum_{i} x_i^{k0}\right) * 10^F + \sum_{i}\sum_{j} d_{ij}y_{ij} \quad (1)$$

Subject to:

$$\sum_{i} x_i^{k0} \leq 1, \forall k \quad (2)$$

$$\sum_{i}\sum_{c} x_i^{kc} = 1, \forall i \quad (3)$$

$$x_i^{kc} \leq \sum_{j \in N_i} x_j^{k(c-1)}, \forall i, k, c \quad (4)$$

$$\sum_{i}\sum_{c} x_i^{kc} l_i \geq T \sum_{i} x_i^{k0}, \forall k \quad (5)$$

$$y_{ij} \geq \sum_{c} x_i^{kc} + \sum_{c} x_j^{kc} - 1, \forall i, j, k \quad (6)$$

$$x_i^{kc} \in \{0,1\}, \forall i, k, c \quad (7)$$

$$y_{ij} \in \{0,1\}, \forall i, j$$

The objective, (1), has two main terms with one term maximizing the number of regions, $\sum_k \sum_i x_i^{k0}$, and the other term minimizing the total within-region dissimilarity, $\sum_i \sum_j d_{ij}y_{ij}$. The number of regions is multiplied by a scaling factor 10^F so that the goal of maximizing the number of regions always dominates the goal of minimizing the total within-region heterogeneity. That is, a solution with larger number of regions will always be preferred over any other solutions with smaller number of regions; for solutions with the same number of regions, a solution with lower heterogeneity will be preferred. Constraints (2) ensure that each region has at most one root unit. Constraints (3) specify that each unit is assigned to exactly one region with one contiguity order. Constraints (4) require that unit i is assigned to region k at contiguity order c if and only if one of its adjacent unit j is assigned to region k at order $c-1$. Constraints (5) ensure that the total value of spatially extensive attribute at each region exceeds the prespecified threshold. Constraints (6) link the decision variables. Constraints (7) impose binary restrictions on decision variables.

While only one spatially extensive attribute was included in this original formulation of max-p-regions, Folch and Speilman (2014) generalized it to enable multiple attributes to be the size constraints for identified regions. Such size constraints combined with the objective of maximizing the number of regions allow for the preservation of as much geographic detail as possible. In addition, the contiguity constraints and the other objective of minimizing the within-region heterogeneity ensure that the identified region is contiguous and as homogeneous as possible. These characteristics make the max-p-regions ideally suited for neighborhood delineation.

However, the max-p-regions is NP-hard and computationally expensive to solve (Duque et al. 2012). The largest-sized problem that can be solved optimally using exact MIP solution method is a problem with 16 units (Duque et al. 2012). To address its associated computational challenges, Duque et al. (2012) developed a two-phase heuristic method with the first phase constructing the feasible solution and the second phase improving the solution from the first phase through several different local search strategies (greedy, simulated annealing, and tabu search). While this heuristic method makes it computationally possible to solve practically sized problems, it takes 10 to 20 hours to get the best quality solutions for problems with over 3,000 units (Duque et al. 2012). There is a clear need to develop more efficient solution approaches for the max-p-regions in order to enable its application to large-scale neighborhood delineation.

Solution approach

Given the computational complexity associated with solving the max-p-regions exactly and heuristically, a new solution approach is developed to efficiently solve max-p-regions for large-sized problems. This new solution approach is composed of three main stages: region growth, enclave assignment, and local search. The first stage focuses on growing regions in such a way that can maximize the number of regions; the second stage assigns enclaves using a randomized greedy strategy; and the final stage iteratively improves the total within-region heterogeneity through a customized simulated annealing that integrates a tabu list. The overall design of the new solution approach for max-p-regions is summarized in Figure 1. After initialization, the procedure of growing regions is repeated for MI times as significant randomness is involved in the procedure and the resulting partition will be different from run to run. Next, the partitions leading to the maximal number of regions are passed to the following procedures for enclave assignment and local search. At the end, the partition having the least within-class heterogeneity is considered to be the best solution identified. Details of the three stages are now presented.

Region growth

The purpose of the region growth phase is to identify a set of contiguous regions whose total spatially extensive attribute exceeds the threshold. The flow chart for region growth is shown in Figure 2. It starts by randomly selecting an unassigned unit as the seed unit for a region and then iteratively adds the unassigned neighbors of the units in the region until it reaches the threshold or no unassigned neighbor can be found. If the region formed fails to reach the threshold, all the units assigned to the region are referred to as “enclave” and are added to the enclave set. This process is repeated until all units have been either assigned to a region or included in the enclave set. At the end of this phase, we will identify a set of contiguous regions whose spatially extensive attribute exceeds the prespecified threshold and a set of enclaves.

This phase focuses on identifying as many regions as possible and does not account for the attribute dissimilarity between units, which are significant design differences from the region growth algorithm proposed in Duque et al. (2012) that grows region by iteratively including the neighboring unit that minimizes the total within-class dissimilarity. As the number of identified regions is determined in this phase and will not be modified in the following two phases, it is important to devise the region growth strategy so that the number of regions can be maximized. The computational results in the next section show that this new algorithm can identify more compact regions and results in much larger number of regions found.

Enclave assignment

The goal of enclave assignment phase is to assign the enclaves to the regions identified in region growth phase. The flow chart for enclave assignment is shown in Figure 3. It starts by randomly selecting a unit in the enclave set and then if any of its neighbors has been assigned to a region, the dissimilarity between the enclave and all neighboring regions are computed and the enclave will be randomly assigned to one of the neighboring regions with the N smallest dissimilarity. This process is repeated until all enclaves have been assigned to a region. At the end of this phase, we will identify a feasible solution for the max-p-regions problem where each region satisfies the contiguity and spatial threshold constraints and the identified regions are a complete partition for the spatial units.

This enclave assignment strategy is different from the greedy enclave assignment in Duque et al. (2012) where each enclave will be assigned to the neighboring region with the smallest dissimilarity. The strategy of randomly choosing one of the best candidates but not necessarily the top candidate is generally referred to as randomized greedy algorithm. It was first introduced by Feo and Resende (1995) in the Greedy Randomized Adaptive Search Procedure (GRASP) to increase solution diversity while not necessarily compromising the solution quality in the initial solution construction. Given such superiority to traditional greedy algorithm, this randomized greedy strategy has been applied in various regionalization problems (Gonzalez-Ramirez et al. 2011; Cano-Belman et al. 2012; Li et al. 2014).

Local search

After identifying a good initial feasible solution in the first two phases, we design a local search algorithm to improve the solution's total within-class heterogeneity by iteratively moving a spatial unit from its current region (donor region) to a neighboring region (recipient region) while ensuring the solution's feasibility. The flow chart for the local search algorithm is depicted in Figure 4. This algorithm follows the general design of simulated annealing (SA) that simulates the process of heating a material and then slowly lowering the temperature to control defect. Duque et al. (2012) has implemented the SA to solve the max-p-regions problem. Specifically, given a feasible solution the SA algorithm identifies all candidate units that can move to a neighboring region without violating the contiguity and threshold constraints, and then randomly selects one candidate unit. If this move can reduce the total heterogeneity, it is accepted; otherwise, the nonimproving is accepted with a probability defined by Boltzmann's equation, $p = e^{-\Delta H/t}$, where ΔH is the total heterogeneity change due to this move and t is the current temperature. This process is iterated with t gradually decreasing at a cooling rate α until t reaches a prespecified value.

Our new algorithm introduces several significant changes to the original SA algorithm. First, our algorithm dynamically updates a list of potential units that can move to a neighboring region without

violating the contiguity and threshold constraints, rather than recompute the potential units at each iteration. Identifying movable units is computationally intensive because for each unit we need check whether losing the unit will break the spatial threshold constraint and whether it will leave the remaining units in the region to be unconnected. Our algorithm recomputes the movable units only when the list of potential units is empty. Otherwise, the list is updated after each move by removing the moved unit, and all the units in the donor and recipient region. This will ensure the rest of units in the list are still feasible to move without violating the constraints. Second, once the potential unit is selected, only the best possible move is considered for further assessment, rather than any possible move. That is, only the neighboring region with the smallest dissimilarity could be the recipient region. As the solution diversity is maintained by randomly selecting candidate unit, allowing best move only could lead to faster convergence to high-quality solution. Third, a tabu list is integrated in the criteria for accepting nonimproving moves. The tabu list that represents a list of banned moves is used in tabu search algorithm to discourage the search from coming back to previously visited solution (Glover 1989). Li et al. (2014) show that once a nonimproving move is made near the algorithm completion, the search bounces among a small set of solutions that consist of reverse moves of previous improving moves. We therefore construct the tabu list by iteratively adding the reverse moves of improving moves to prevent this and result in faster convergence. A nonimproving move is made only when it is not in the tabu list and the Boltzmann's probability is larger than a random value. The tabu list has a prespeficied length limiting the number of moves that can be accommodated in the list, and takes the queue strategy when the list is full. Finally, our algorithm allows for termination when all of the previous NC potential moves selected are nonimproving, rather than only in the case where the temperature t reaches a predefined value. This termination condition is consistent with the condition for tabu search in Duque et al. (2012). Computational experiments show this termination condition could lead to better-quality solutions.

In addition to the SA, Duque et al. (2012) also tested tabu search and greedy algorithms for local search. They reported that the tabu search can identify the best solutions in most scenarios but it is much more computationally expensive, whereas the simulated annealing and greedy algorithms are computationally efficient but lack the capacity to identify the best solutions. This new local search algorithm combines the strengths of tabu search and simulated annealing with the aim of identifying better-quality solutions and improving computational efficiency.

Results

We performed a series of computational experiments to assess the performance of the proposed approach for solving the max-p-regions problem. The data sets are retrieved from sample data in the ClusterPy library for regionalization research (Duque et al. 2011). The data include four simulated data sets, which are regular lattices with 100, 529, 1,024, 2,025 units, and two realistic datasets, which are 58 counties in California and 3106 connected counties in the U.S. The attribute value to measure the dissimilarity d_{ij} for the regular lattices is simulated using a spatial autoregressive process with $p = 0.9$, whereas the spatial extensive attribute value l_i is simulated using a uniform distribution of [10, 15]. Three different threshold values $T = 100$, 300, and 500 are tested for the simulated data set. The attribute dissimilarity d_{ij} is also simulated for the counties in California but median household income is used for the counties in the U.S. The spatial extensive attribute value l_i is the population for the counties in California and the number of household units for the counties in the U.S. Three different threshold values $T = 100{,}000$, 300,000, and 500,000 are tested for the two realistic data set.

As the number of identified regions is determined in the region growth phase, we first run the new region growth algorithm 999 times for each combination of the dataset and threshold to compare the number of regions with what is found using the region growth approach in Duque et al. (2012). The results are reported in Figure 5, which shows that our new region growth algorithm identified larger number of regions for all datasets and thresholds except the dataset of counties in California. This is probably because of its small number of spatial units. For example, the number of regions identified by our new algorithm for the 2,025 unit regular lattice with $T = 100$ ranges from 198 to 213 during the 999 runs, whereas that by the approach in Duque et al. (2012) ranges from 187 to 205. For the U.S. counties dataset with $T = 500{,}000$, the number of regions identified by the new algorithm varies from 148 to 165 during the 999 runs, whereas that by the approach in Duque et al. (2012) ranges from 137 to 154. Clearly, our new algorithm generally dominates the approach in Duque et al. (2012) in terms of number of regions identified.

Next, for each partition with the maximum number of regions, we assign enclave using our new algorithm to generate initial feasible solutions. While several different local search algorithms are used in Duque et al. (2012), tabu search generally identified the best quality solutions. As a result, we only compare our local search algorithm with the tabu search in Duque et al. (2012). In order to make the results comparable, we run our local search algorithm and tabu search with the same feasible solution generated in previous stages. Each of the local search algorithms is run 10 times and the best solution is reported. The computational results are reported in Table 1. The column "Total heterogeneity reduction" is defined as:

$$Total\ heterogeneity\ reduction = \frac{h(inital\ solution) - h(final\ solution)}{h(inital\ solution)} \qquad (8)$$

where h represents the total within-class heterogeneity to evaluate the improvement of total within-class heterogeneity by local search algorithms. For datasets lattice 100, lattice 529, and California counties, our new local search algorithm leads to an average of 10.92%, 1.47%, 11.12% more total heterogeneity reduction for all three thresholds, respectively. For lattice 1024, our new local search algorithm results in 2.19% and 3.34% more total heterogeneity reduction for $T = 100$ and 500, respectively. For lattice 2025, it performs 0.18% and 0.39 % better for $T = 300$ and 500, respectively. For US counties, tabu search performs better for all three thresholds with 0.72%, 1.62% and 0.55% more total heterogeneity reduction. Column "Running time" reports the computational time to run the local search algorithm. The tabu search takes more time in all scenarios except one for lattice 100 and one for California counties. The speedup of our new local search algorithm compared with tabu search is significant for larger data sets. For example, the speedup for lattice 2025 ranges from 12 to 118 and for US counties it ranges from 22 to 92.

Table 1: Computational results of new local search algorithm and tabu search algorithm

Dataset	Threshold	Total heterogeneity reduction		Running time (seconds)	
		New algorithm	Tabu search	New algorithm	Tabu search
Lattice 100	100	30.99%	13.12%	0.24	2.67
Lattice 100	300	12.99%	11.37%	0.17	0.14
Lattice 100	500	32.32%	19.04%	0.83	4.66
Lattice 529	100	30.10%	29.66%	2.69	124.24
Lattice 529	300	23.17%	22.31%	3.09	17.50

Lattice 529	500	28.24%	25.12%	6.79	27.01
Lattice 1024	100	24.58%	22.40%	5.66	53.80
Lattice 1024	300	25.08%	26.87%	11.18	67.35
Lattice 1024	500	21.17%	17.82%	7.71	36.77
Lattice 2025	100	28.00%	28.46%	13.27	1560.19
Lattice 2025	300	24.12%	23.94%	20.39	240.00
Lattice 2025	500	25.84%	25.45%	28.90	1262.35
CA counties	100,000	17.82%	2.59%	0.09	0.04
CA counties	300,000	36.78%	29.79%	0.13	0.13
CA counties	500,000	42.64%	31.52%	0.11	0.88
US counties	100,000	28.54%	29.26%	39.66	3641.65
US counties	300,000	27.99%	29.62%	39.70	887.40
US counties	500,000	21.30%	21.85%	72.38	3383.63

Discussion and Conclusions

The last several decades have borne witness to three important trends in urban social science. The first--rapidly expanding data resources--is not limited to the urban context. Indeed, in recent years exploding volumes of data have led to the rapid development of techniques for both Big Data analysis and the data pipelining process. In urban research, however, this trend is also accompanied by (1) an increasing topical focus on neighborhoods and the important roles they play in human development and global sustainability, and (2) an increasing awareness of linked and multilevel spatial processes and the development of analytical techniques used to study them (Raudenbush & Bryk, 2002; Raudenbush 2003; Harris 2007; She, Duque, & Ye, 2017; Zhong et al 2019;). In practice, these trends mean that the *problem size* in quantitative geography is increasing by orders of magnitude. Put differently, researchers today seek answers to questions about multiscalar neighborhood growth and change, persistent neighborhood inequality in high-performing economies, or neighborhood processes that link together places, actors and institutions within a single modeling framework. Addressing these challenges requires not only increasingly powerful computational platforms but also more efficient and performant implementations of the fundamental algorithms for urban neighborhood research. In this paper, we present one such advance.

The Max-p-regions algorithm is designed to partition a study area into the largest possible set of mutually exclusive regions (or neighborhoods) that still satisfy an internal homogeneity constraint. Since its inception in 2012 (Duque et al. 2012a), the max-p-regions has been applied in various urban and social contexts including urban slum delineation (Duque et al. 2012b), neighborhood dynamics (Rey et al. 2011), urban energy assessment (Reyna et al. 2016), regional inequality analysis, and more recently has been extended to problems that address network connectivity (She, Duque, & Ye 2017) and interregional comparisons (Rey and Sastré-Gutiérrez, 2010). Despite the important findings advanced by these studies, we argue that the current implementation is waning in utility, since it is unable to accommodate the massive data requirements inherent in modern urban scholarship.

In this paper, we have developed a new solution algorithm for max-p that can substantially reduce its computation time, and thus facilitates a much broader set of use-cases and larger volume of input data. This means, for instance, that scholars are now able to leverage max-p to address metropolitan-scale comparative research in hours or days that would previously take weeks or months. Beyond its

substantial improvement in runtime, however, our new algorithm also improves solution quality substantially by identifying much larger number of regions that also realize smaller within-region heterogeneity in comparison with the original algorithm in Duque et al. (2012).

References

Anselin, L. (1995). Local indicators of spatial association-LISA. Geographical Analysis, 27(2), 93–115.

Cano-Acosta, A., Fontecha, J., Velasco, N., & Muñoz-Giraldo, F. (2016). Shortest path algorithm for optimal sectioning of hydrocarbon transport pipeline. IFAC-PapersOnLine, 49(12), 532-537.

Clark, W. A. V., Anderson, E., Östh, J., Malmberg, B., Osth, J., & Malmberg, B. (2015). A Multiscalar Analysis of Neighborhood Composition in Los Angeles, 2000-2010: A Location-Based Approach to Segregation and Diversity. Annals of the Association of American Geographers, 105(6), 1260–1284.

Delmelle, E. C. (2016). Mapping the DNA of urban neighborhoods: Clustering longitudinal sequences of neighborhood socioeconomic change. Annals of the American Association of Geographers, 106(1), 36–56.

Duque, J. C., Anselin, L., & Rey, S. J. (2012). The max-p-regions problem. Journal of Regional Science, 53(3), 397–419. https://doi.org/10.1111/j.1467-9787.2011.00743.x

Duque, J. C., Church, R. L., & Middleton, R. S. (2011). The p-Regions Problem. Geographical Analysis, 43(1), 104–126. https://doi.org/10.1111/j.1538-4632.2010.00810.x

Duque, J. C., Ramos, R., & Surinach, J. (2007). Supervised Regionalization Methods: A Survey. International Regional Science Review, 30(3), 195.

Feo, T. A., & Resende, M. G. (1995). Greedy randomized adaptive search procedures. Journal of global optimization, 6(2), 109-133.

Folch, D. C., & Spielman, S. E. (2014). Identifying regions based on flexible user-defined constraints. International Journal of Geographical Information Science, 28(1), 164–184. https://doi.org/10.1080/13658816.2013.848986

Garreton, M., & Sánchez, R. (2016). Identifying an optimal analysis level in multiscalar regionalization: A study case of social distress in Greater Santiago. Computers, Environment and Urban Systems, 56, 14-24.

Glover, F. (1989). Tabu search—part I. ORSA Journal on computing, 1(3), 190-206.

González-Ramírez, R. G., Smith, N. R., Askin, R. G., Miranda, P. A., & Sánchez, J. M. (2011). A hybrid metaheuristic approach to optimize the districting design of a parcel company. *Journal of Applied Research and Technology*, *9*(1), 19-35.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). International Journal of Geographical Information Science, 22(7), 801–823. https://doi.org/10.1080/13658810701674970

Guo, D. and Wang, H. (2011), Automatic Region Building for Spatial Analysis. Transactions in GIS, 15: 29-45. doi:10.1111/j.1467-9671.2011.01269.x

Gurobi Optimization. (2019). Inc.,“Gurobi optimizer reference manual,” 2019.

GNU Linear Programming Kit. (2012). GLPK. "https://www.gnu.org/software/glpk/"

Harris, R., Johnston, R., & Burgess, S. (2007). Neighborhoods, Ethnicity and School Choice: Developing a Statistical Framework for Geodemographic Analysis. Population Research and Policy Review, 26(5–6), 553–579. https://doi.org/10.1007/s11113-007-9042-9

Harris, R., Sleight, P., & Webber, R. (2005). *Geodemographics, GIS and neighbourhood targeting* (Vol. 8). John Wiley & Sons.

Kim, H., Chun, Y., & Kim, K. (2015). Delimitation of Functional Regions Using ap-Regions Problem Approach. *International Regional Science Review*, *38*(3), 235-263.

Li, W., Church, R. L., & Goodchild, M. F. (2014). The p-compact-regions problem. *Geographical Analysis*, *46*(3), 250-273.

Openshaw, S., & Rao, L. (1995). Algorithms for reengineering 1991 Census geography. Environment and planning A, 27(3), 425-446.

Openshaw, S. P., T.(1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *Statistical Applications in the Spatial Sciences*, 127-144.

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

Raudenbush, S. W. (2003). The Quantitative Assessment of Neighborhood Social Environments. In Neighborhoods and Health (pp. 112–131). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195138382.003.0005

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (Vol. 1). Sage.

Rey, S. J., Anselin, L., Folch, D. C., Arribas-Bel, D., Sastré Gutiérrez, M. L., & Interlante, L. (2011). Measuring spatial dynamics in metropolitan areas. *Economic Development Quarterly*, *25*(1), 54-64.

Rey, S. J., & Sastré-Gutiérrez, M. L. (2010). Interregional inequality dynamics in Mexico. Spatial Economic Analysis, 5(3), 277-298.

Reyna, J. L., Chester, M. V., & Rey, S. J. (2016). Defining geographical boundaries with social and technical variables to improve urban energy assessments. *Energy*, *112*, 742-754.

She, B., Duque, J. C., & Ye, X. (2017). The Network-Max-P-Regions model. International Journal of Geographical Information Science, 31(5), 962–981. https://doi.org/10.1080/13658816.2016.1252987

Spielman, S. E., & Singleton, A. (2015). Studying Neighborhoods Using Uncertain Data from the American Community Survey: A Contextual Approach. Annals of the Association of American Geographers, 105(5), 1003–1025.

Zhong, Q., Karner, A., Kuby, M., & Golub, A. (2019). A multiobjective optimization model for locating affordable housing investments while maximizing accessibility to jobs by public transportation. Environment and Planning B: Urban Analytics and City Science, 46(3), 490–510. https://doi.org/10.1177/2399808317719708

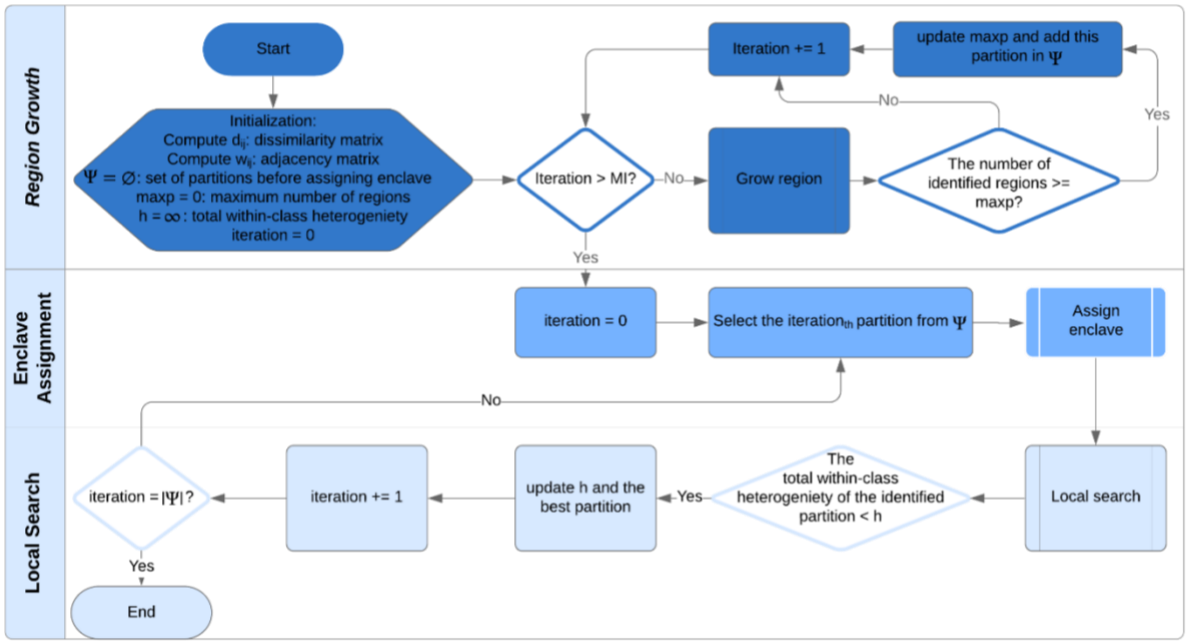

Figure 1: Flow chart of the new solution approach

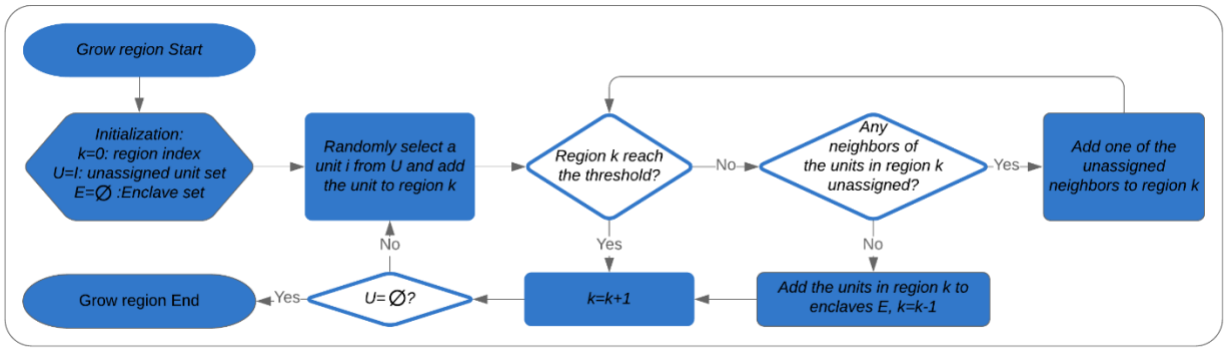

Figure 2: Flow chart of region growth

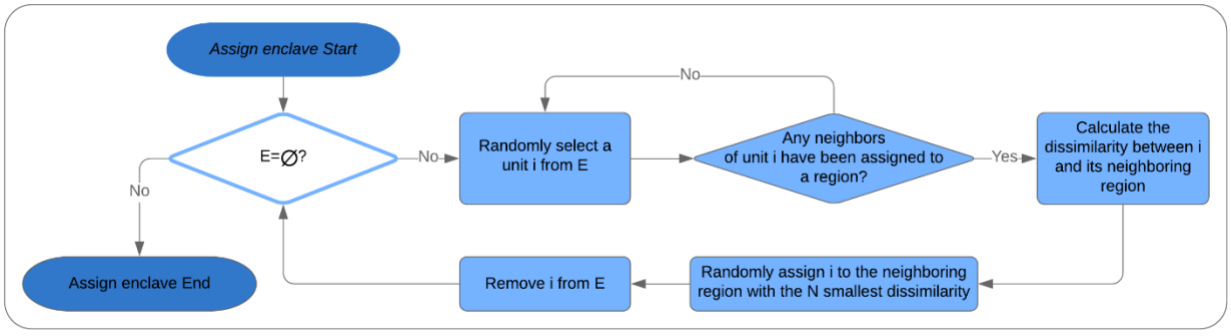

Figure 3: Flow chart of enclave assignment

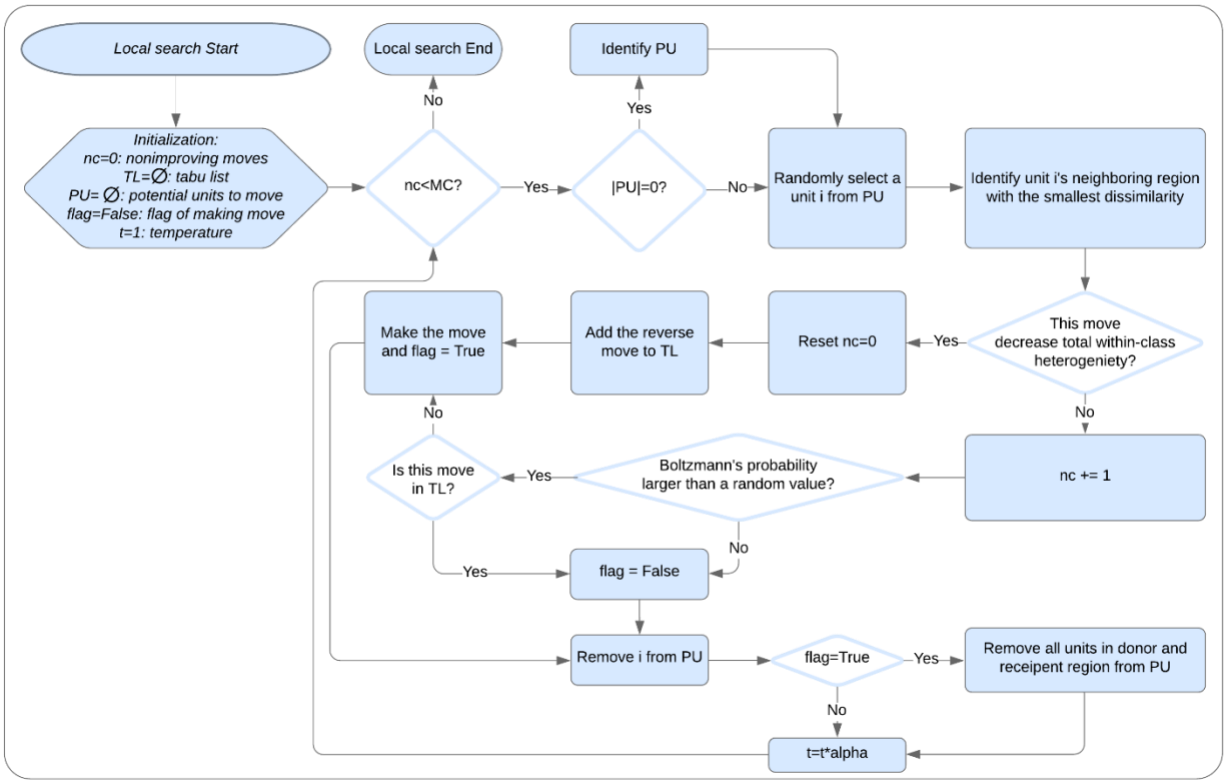

Figure 4: Flow chart of local search

Figure 5: Distribution of the number of identified regions by the new region growth algorithm and the algorithm in Duque et al. (2012)