An AZP-ACO Method for Region-Building

Angelos Mimis*, Antonis Rovolis, and Marianthi Stamou

Panteion University, Department of Economic and Regional Development, Athens, Greece
{mimis,rovolis}@panteion.gr,
marianthi_stamou@hotmail.com

Abstract. In this paper a regionalization algorithm which groups spatial areal objects into homogeneous zones is presented. The proposed method is based on Automatic Zoning Problem (AZP) procedure which is extended to use the Ant Colony Optimization (ACO) technique. The results produced are compared to the original AZP method. Both methods are applied into the classification of economic data in a post code level on the area of Athens.

Keywords: Region building, constrained clustering, Ant colony optimization, Automatic Zoning Problem.

1 Introduction

The zone design problem involves the aggregation of k regions into n zones while optimizing an objective function and preserving the internal connectivity of the zones [1-2]. This problem is also known as redistricting, regionalization or p-region problem, and in terms of computational complexity, it belongs to the family of nondeterministic polynomial-time hard problems (N-P hard) [3-4].

The zone design is a geographical problem that has been applied to many fields such as climate zoning [5], location optimization [6], in socio-economic [2], [7] and epidemiological analysis [8], in electoral and school districting [4], [9] and many more. The problems mentioned may differ into the data types (numerical or categorical) the different objective functions used (e.g. capturing intra-region homogeneity or compactness) or the constraints imposed (e.g. minimum population within zones). All of these approaches can be classified into the following three categories, 1) linear programming techniques, 2) heuristic-based optimization and 3) contiguity constrained clustering [10-12].

In the first group of methods, Duque et al [12-13] have formulated the regionalization as a mixed integer programming problem and incorporated the contiguity within zones in the solution space searched. This approach is computational expensive and thus its use is limited to small data sets.

In the second group, the heuristic-based optimization techniques are included, which optimize a given function while preserving the contiguity constraints. The first method developed was the Automatic Zoning Problem (AZP) by Openshaw [1] which

* Corresponding author.

I. Maglogiannis, V. Plagianakos, and I. Vlahavas (Eds.): SETN 2012, LNAI 7297, pp. 81–89, 2012.

starts with an initial solution and by randomly adding neighboring regions to various zones, converges to a better solution. This approach was later improved by applying Tabu search and Simulating Annealing in order to avoid the local optimum in the search space (ZDES software) [2]. In a more recent approach by Bacao et al. [4], a Genetic Algorithms was developed and compared with the ZDES results. These methods have two inherited limitations. Firstly, they are computational intensive and secondly, some variants give significantly different results with each run.

The third approach is based on hierarchical clustering and its more recent form is performed into two steps. Initially a hierarchical constrained clustering is performed to create the tree of aggregated regions, followed by an optimization method in order to determine the cut of level in the hierarchical tree [7], [10-11]. In the hierarchical step various methods have been implemented such as the single, average, complete linkage and Ward method. The drawback of this methodology is the limited solution space that is explored.

In this paper the second approach is adopted, aiming to an improvement to the current methodology in use. The AZP method is extended to use the Ant Colony Optimization (ACO) technique and the results produced are compared to the initial AZP method. The advantages of our method are threefold: to avoid local optimum, to provide improved solutions and a tool that can be used alongside the Geographical Information System (GIS) software.

In the next section the proposed methodology is presented. In section 3 the algorithm is applied on a case study on the city of Athens, where the data and results are presented. Finally in the last section some concluding remarks are drawn.

2 Methodology

In the problem at hand, a study area which is completely divided into regions is given. These regions are aggregated into zones so that each region is assigned to only one zone and that the zones are contiguous internally (regions within a zone) and externally (zones together). In order to measure the quality of a partition an objective function of the attribute values of the regions for the current zone-design is examined, aiming in optimum performance.

In the initial work of Openshaw [1], a mildly steepest descent algorithm was proposed (AZP). This algorithm consists by the following steps:

- Step 1. An initial solution is found, or in terms of the problem an initial classification of n regions into a given number of m zones (m<n).
- Step 2. A list of the zones $L=\{Z_1,Z_2,\ldots,Z_m\}$ is made.
- Step 3. A zone Z_k is randomly selected and removed from L.
- Step 4. For Z_k a list of contiguous regions $B=\{R_1,R_2,\ldots\}$ is composed.
- Step 5. Until this set of regions B is empty, a region R_j is selected randomly and removed from B.
- Step 6. The zone in which R_j belongs is found and it is examined if it can be removed from it. If by aggregating R_j into zone Z_k the zone remains contiguous then the algorithm continues to step 7, otherwise to step 5.

- Step 7. The value of the objective function is calculated for the new classification.
- Step 8. If an improvement is made, the new zonation is kept, otherwise it is rejected and the algorithm moves to step 5.
- Step 9. Set B is re-composed and the algorithm moves to step 5.

The steps 2-9 are repeated until convergence is achieved.
ACO is inspired by nature, where ants select the shortest path to food by laying and following chemical trails called phenomone. After the initial paper by Dorigo and Cambardella [14], it has been applied to various problems except Travelling Salesman Problem (TSP) such as vehicle routing, graph coloring, sequential ordering [15-16]. The algorithm is performed by a number of iterations. In each iteration, in a typical ACO application, the steps are the following:

- Step 1. A set of m ants are located at randomly selected regions.
- Step 2. Each ant is making a tour through the neighboring regions, visiting each region once.
- Step 3. For each ant, say k, the next region q to be visited is selected, based on the probability

$$P_{qk} = \frac{\tau_{zk}^{a} * V_{zk}^{b}}{\sum_{q} \tau_{zk}^{a} * V_{zk}^{b}} \quad (1)$$

  where τ is the phenomone strength, v is the visibility and a, b are two constants. If a=0, the closest region in attribute values is chosen. Also in the case where b=0 the visibility between regions is ignored.

- Step 4. When the ants has explored the space, the phenomone trails are updated by:

$$\tau_{qz} = (1 - \rho)\tau_{qz} + \Delta\tau_{qz} \quad (2)$$

  where ρ is the evaporation constant, and $\Delta\tau_{qz}$ is the reinforcement achieved by region inclusion in zone.

In our version of AZP, called AZP-ACO, the steps 5-6, are replaced by the ACO algorithm described. The initial values, for pheromone, are evaluated in the first step of the algorithm, by examining the difference between the attribute values of contiguous regions. Typical values for the AZP-ACO method are a=b=1, ρ=0.5 and m is equal to the number of neighboring regions.
As an objective function a homogeneity measure is used, which is defined as the sum of squared differences between attributes allocated on each region and the mean values of each zone. The Sum of Squared Differences (SSD) is given by:

$$SSD = \sum_{L} \sum_{i} \sum_{j} (x_{ij} - \overline{x}_{j})^2 \quad (3)$$

where $\overline{x}_j$ is the mean value of j attribute on a zone in L set and x_{ij} is the value of j attribute on region i, classified on a zone in L [7], [11]. The SSD is a measure of dispersion of attribute values for the regions in a zone. Also homogeneous zones contribute small values into the objective function.

The algorithms of AZP and AZP-ACO were implemented in C++ and were loosely coupled with ArcGIS 9.3. In ArcGIS, which is the main tool in our analysis, the regions are defined and are accompanied by the relative attribute data. In ArcGIS the W contiguity matrix is evaluated and exported with the attribute values into the C++ program. In our approach the W matrix has the values of 1 for regions with common part of an edge and 0 otherwise (rook contiguity). Finally the output of the custom program, which is the regions classification, is passed into the GIS software to display and further examine the result.

3 Data

Our empirical analysis is based on the postcode on the city of Athens. More specifically, the area of our analysis contains of 207 postcodes (Figure 1). The variable considered for classification is the average annual income, published by the Greek Ministry of Finance in 2002 (Available at http://www.gsis.gr/ggps/statistika/statistika.html).

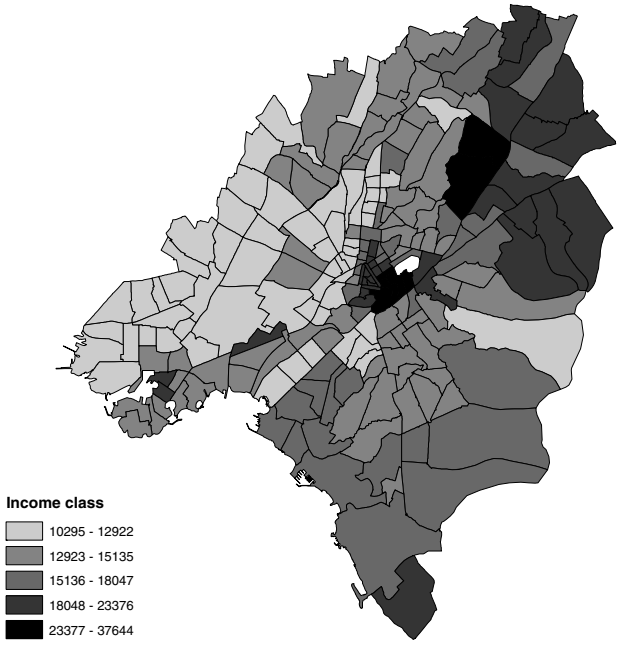

Fig. 1. Study area

4 Results

The AZP and the AZP-ACO were applied to the regionalization of the post codes in the study area of Athens. The spatial objects, in our case, have only one attribute, the mean income.

The methods have been compared for speed and quality of the solution. Four experiments were made, in which the number of spatial objects are the same (207) as well as the number of attributes. In the first case, the resulting clusters were 5 and in the other experiments were 10, 15 and 20 respectively. The experiments were run 30 times for each method and the average results are presented in Table 1, where the number of function evaluations, the time needed and the value of the objective function are presented.

Table 1. Results of the AZP and AZP-ACO algorithm

	AZP			AZP-ACO		
#regions	#evaluations	time(sec)	value	#evaluations	time(sec)	value
5	437	10.8	14.9	7310	140.6	10.9
10	559	8.4	11.5	6843	112.6	7.5
15	645	7.2	10.0	6660	102.5	5.7
20	721	6.8	8.9	5987	86.4	4.5

As far as the quality criterion is concerned, the internal homogeneity of the resulting zones was measured by Equation (3), where smaller values are better. In all the tests the AZP-ACO produces improved results. As it can be seen from the Table 1, as the number of regions increases, the ratio $SSD_{AZP}/SSD_{AZP-ACO}$ increases as well, starting from 1.37 for 5 regions and ending up to 1.97 for 20 regions. On the other hand, the time needed to perform the clustering is around 13 times more for AZP-ACO than for the AZP. Further, as can be seen in Table 2, AZP-ACO is superior to classic AZP not only in the mean value of the objective function but in the standard deviation as well.

Another important issue that is usually ignored in the literature is to visualize the zoning proposed by the methods and compare the results in terms of geography. For that purpose in mind a typical zoning for the two algorithms is presented in Figures 2-5. In the first case, regions are aggregated into 5 zones. The difference in Figures 2 and 3 is substantial. The AZP-ACO method (Figure 3) has captured the areal changes in income and has divided the region, as expected, into the city center area (zone 1), the surrounding of the city center (zone 3), the north-west (zone 5), the south-east (zone 4) and the north-east zones (zone 2). On the other hand AZP, did not distinguish the north-east zone from south-east zone (zone 3 in Figure 2) and has displayed a limited area surrounding the city center (zone 2). In the second case, 10 zones were used and both methods have captured the characteristics of the data set. As can be seen in Figures 4 and 5 both methods have created zones for the high, medium and low income regions (Figure 1). A striking difference in the two zoning-designs lies in zone 6 in Figure 4, which by looking into the data it can be seen, that AZP has included non-homogeneous regions. A more robust approach has been adopted by

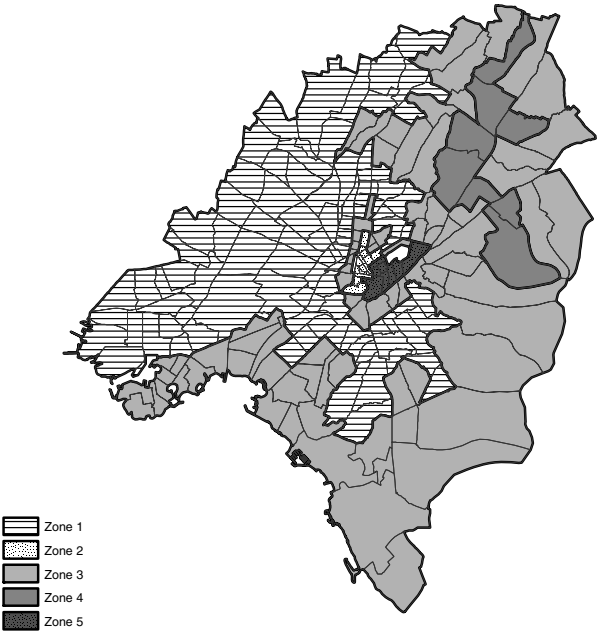

Fig. 2. The region-design proposed by AZP for 5 regions

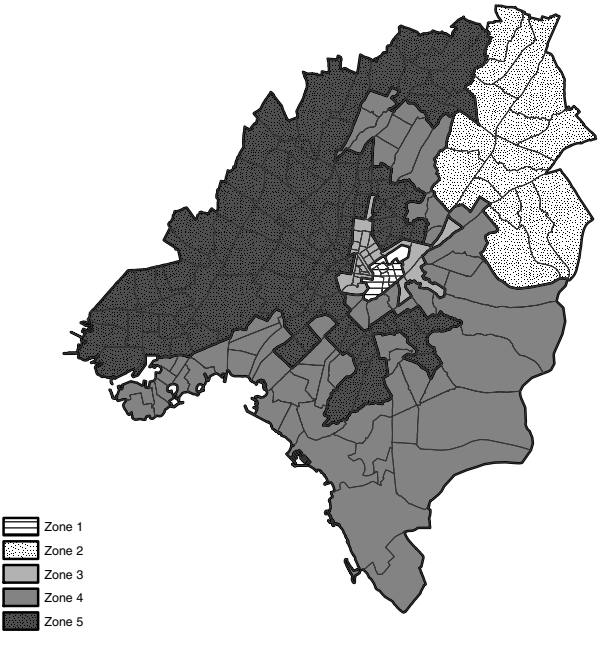

Fig. 3. The region-design proposed by AZP-ACO for 5 regions

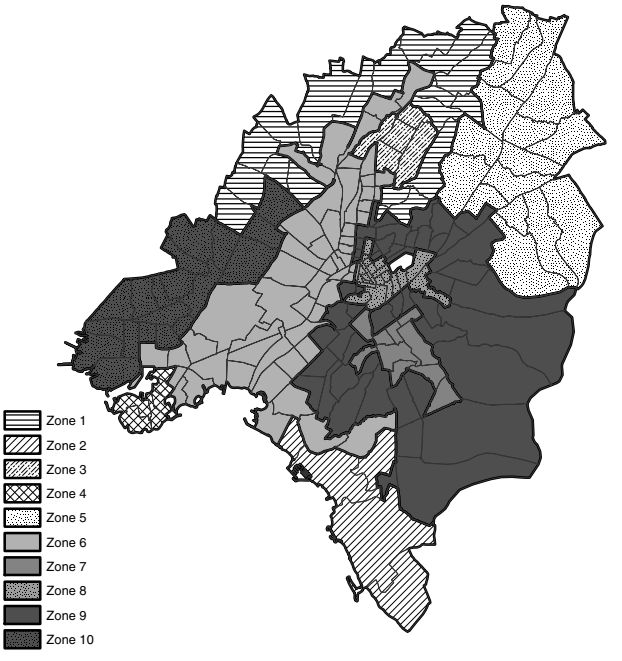

Fig. 4. The region-design proposed by AZP for 10 regions

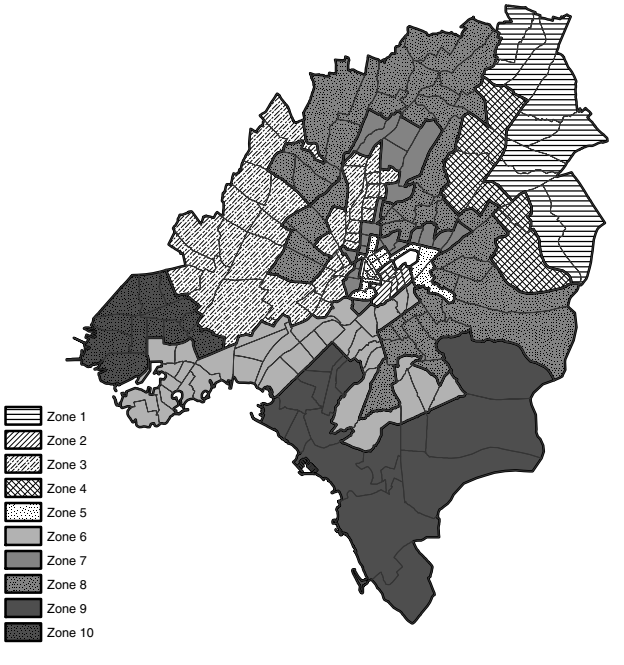

Fig. 5. The region-design proposed by AZP-ACO for 10 regions

AZP-ACO (Figure 5) where the same area has been partition and aggregated into zones 6 and 3 resulting in a more natural partitioning. Further, a drawback of the illustrated clustering, for both cases, is the presence of long zones (e.g. zone 1 in Figure 4 and zone 8 in Figure 5). This is attributed to the data set used (Figure 1) and can be treated by including into the objective function a term for compactness of the zones.

Table 2. Descriptive statistics of the results

	AZP			AZP-ACO		
#regions	median	mean	stdev	median	mean	stdev
5	14.5	14.9	3.8	9.8	10.9	3.1
10	11.0	11.5	3.2	7.5	7.5	2.1
15	9.6	10.0	2.7	5.5	5.7	1.3
20	8.5	8.9	2.3	4.3	4.5	1.1

Finally, an issue commonly arising in region clustering is the number of zones that should be used. As Guo [10] pointed out, for explanatory spatial data analysis, the number of zones is a problem parameter and the results are examined for various values. On the other hand, it can be seen that for particular applications, the problem definitions depicts the number of zones as in the case of electoral or school districting [4].

5 Conclusions

In this paper, an efficient method for regionalization is presented, which combines the original AZP methodology with the ACO method of optimization. This approach was compared with the original method of Openshaw and illustrated an improvement of more than 30% in terms of the objective function.

The region-design method proposed can be applied into different domains and permits the use of different definitions for contiguity and objective function in terms of compactness, dissimilarity or heterogeneity.

References

1. Openshaw, S.: A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modeling. Trans. Inst. Br. Geogr. 2, 459–472 (1977)
2. Openshaw, S., Rao, L.: Algorithms for reengineering 1991 census geography. Environ. Plann. A 27, 425–446 (1995)
3. Keane, M.: The size of the region-building problem. Environ. Plann. A 7, 575–577 (1975)
4. Bacao, F., Lobo, V., Painho, M.: Applying genetic algorithms to zone design. Soft. Comput. 9, 341–348 (2005)
5. Fovel, R.G., Fovell, M.-Y.C.: Climate zones of the conterminous United States defined using cluster analysis. J. Clim. 2, 2103–2135 (1993)
6. Goodchild, M.F.: The aggregated problem in location allocation. Geogr. Anal. 11, 240–255 (1979)

7. Assuncao, R.M., Neves, M.C., Gamara, G., Da Costa Freitas, C.: Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. Int. J. Geogr. Inform. Syst. 20(7), 797–811 (2006)
8. Haining, R., Wise, S., Blake, M.: Constructing regions for small-area analysis material deprivation and colorectal cancer. J. Public Health Med. 16, 457–469 (1994)
9. Martin, D.: Automated zone design in GIS. In: Atkinson, P., Martin, D. (eds.) GIS and Geocomputation, Innovations in GIS, vol. 7, pp. 103–111. Taylor and Francis, London (2000)
10. Guo, D.: Regionilization with dynamically constrained agglomeration clustering and partitioning. Int. J. Geogr. Inform. Sci. 22(7), 801–823 (2008)
11. Guo, D., Wand, H.: Automatic region building for spatial analysis. Transactions in GIS 15, 29–45 (2011)
12. Duque, J.C., Ramos, R., Surinach, J.: Supervised regionalization methods: A survey. Int. Reg. Sci. Rev. 30, 195–220 (2007)
13. Duque, J.C., Church, R.L., Middleton, R.S.: The p-regions problem. Geogr. Anal. 43, 104–126 (2011)
14. Dorigo, M., Gambardella, L.M.: Ant colonies for the travelling salesman problem. BioSystems 43, 73–81 (1997)
15. Bonabeau, E., Dorigo, M., Theraulaz, G.: Inspiration for optimization from social insect behavior. Nature 406, 39–42 (2000)
16. Dorigo, M., Stutzle, T.: Ant colony optimization. The MIT Press, Cambridge (2004)