IDENTIFYING VECTOR FEATURE TEXTURES USING FUZZY SETS

C. Anderson-Tarver [a], S. Leyk [a], B. Buttenfield [a]

[a] Dept. of Geography, University of Colorado-Boulder, Guggenheim 110, 260 UCB Boulder, Colorado 80309-0260, USA - (anderscn, stefan.leyk, babs)@colorado.edu

KEY WORDS: Area-Patch, Fuzzy Set Theory, Automated Model Generalization, Archipelagos, Pattern Recognition

ABSTRACT:

This paper describes a method to approach area-patch problems in model generalization using fuzzy set theory. The area patch problem identifies problems in generalizing areas (polygons) with the same semantics (feature code) but varying geometry and spatial distribution. Area-patch generalization can be understood as a pattern recognition problem with a variety of viable solutions, constrained by the purposes of the generalized output. In this research constraints are defined by common USGS cartographic measures for feature generalization. The vagueness, which is inherent in area-patch problems, stems from special cases such as archipelagos. Archipelagos are collections of polygons, any one of which may not contribute significantly to the pattern of the database, but may become prominent when conceptualized as a group.

The paper demonstrates how to access the vague concept of archipelagos in a GIS environment using fuzzy sets to improve area patch generalization. We develop our method generalizing swamps and marshes in an NHD High-Resolution subbasin dataset spanning the Florida-Georgia border. Fuzzy membership functions are assigned for area, inter-polygon distance and number of neighbors within a predefined distance as known contributors to texture. These attributes are combined in a fuzzy overlay to derive degrees of memberships of polygons to the concept of a prototypical archipelago. The final delineation of archipelagos is based on 'alpha cuts' thresholding. A sensitivity analysis evaluates the impact of alpha-cuts on the resulting pattern recognition.

For validation we compare the change in geometric properties (area, area/perimeter ratios) of polygons and overall texture from the original scale to the target scale between our approach and a solution that does not take into account archipelagos. Preliminary findings indicate that a fuzzy set approach allows for the capture of archipelagos, which would otherwise not be included in a generalization solution.

1. INTRODUCTION

1.1 Fuzzy Logic and Generalization

With the introduction of Fuzzy Set Theory by Zadeh (1965) and its revolutionary concept that classes of objects could have vague membership attributes, there has been much discussion in recent years about the prospects of fuzzy logic for various GIS applications (Fisher 1992, Ahlqvist et al. 2003, Deng and Wilson 2008). While the debate over the intrinsic vagueness of geographic objects and their boundaries continues to the present day, fuzzy logic has proven to be a useful conceptual tool for GIScience researchers who want to analyze uncertainties, understand the limits of their conceptual models, and offer alternative viewpoints for often slightly disingenuous geographic delineations.

One area of GIS that has offered no known examples of a fuzzy logic approach to a geographic problem is generalization. Generalization in the geographic sense is the process of taking information, modifying its complexity, and –typically- displaying it in some sort of map format to convey an overall picture of the source information. In a purely visual context, generalization as a form of abstraction refers to the discrepancy of a display from photorealism. Generalization, is also a type of information processing used to tease out the knowledge deemed important for a given purpose (Brassel and Weibel 1988, Buttenfield and Mark 1991). Important in generalization is the selection of crucial characteristics to display geographic essences, which are always dependent on a purpose (Brassel and Weibel 1988). Since generalization is typically performed for use in a map design, Figure 1 illustrates three interrelated components of map design with generalization, symbolization, and production being the results of the interplay of abstraction and constraints (Buttenfield and Mark 1991).

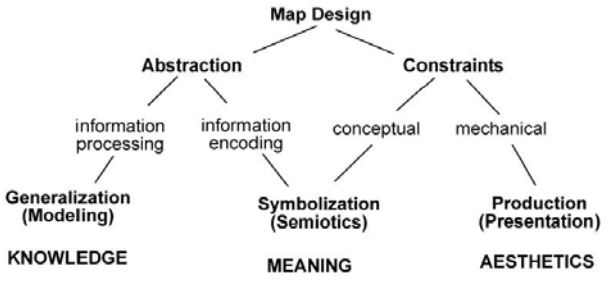


Figure 1: Design Process (from Buttenfield and Mark 1991)

Merging vector features has been and continues to be a difficult task for automated generalization systems (Müller et al. 1995, Bundy et al. 1995, Steiniger and Weibel 2007). In fact, several authors use the generalization problem of polygon aggregation as a case in point to illustrate hurdles automatic generalization

needs to overcome to be more holistic (Bertin 1983, Beard 1991, Steiniger and Weibel 2007). In broad terms the reason why aggregating polygons is so illusive for automated processes is that in many ways computers still have difficulties with pattern recognition if made up for human cognition/ cartographic purposes.
Moreover, many geographic concepts themselves are vague as they relate to the conceptual constraints of map design.
Therefore a fuzzy logic approach to automated generalization in relation to certain vector aggregation problems is not only appropriate but advantageous in order to reconcile ambiguous feature patterns.

1.2 The Area-Patch Problem

The area-patch problem as first identified by Bertin 1983 studies how to manually generalize clusters of marks on a "two dimensional continuum". Originally Bertin (1983) was examining how best to preserve the structural pattern of the lacustrine area of Dombes, France (Figure 2) and thus to explain the spatial relations between data items. Figure 2 describes an examplaric solution to preserve the spatial structure of a dataset given in Bertin (1983). There are only two published automated solutions for the area-patch problem. Müller and Wang (1992) described an algorithmic stepwise solution to a different dataset than the one first illustrated by Bertin (1983). Ormsby and Mackaness (1999) utilize an object-oriented phenomenological paradigm to improve the results of Müller and Wang (1992) and allow for a wider range of feature classes. Since meso and macro-structures of polygon distributions such as islands or lakes exist pole to pole (Ruas 2000), and neither previous study uses a real world dataset, an area-patch dataset representing a corporeal class of features is helpful to better understand the area-patch problem.

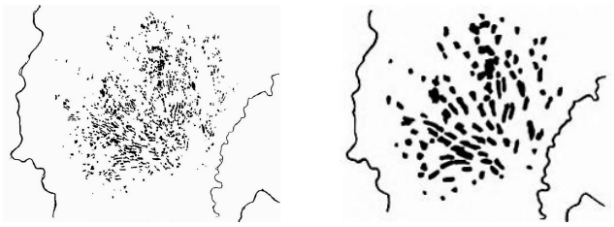


Figure 2. (Left) Area-patch example first studied by Bertin (1983). (Right) Generalization solution proposed by Bertin (1983) after a twenty-fold scale reduction.

1.3 Archipelagos

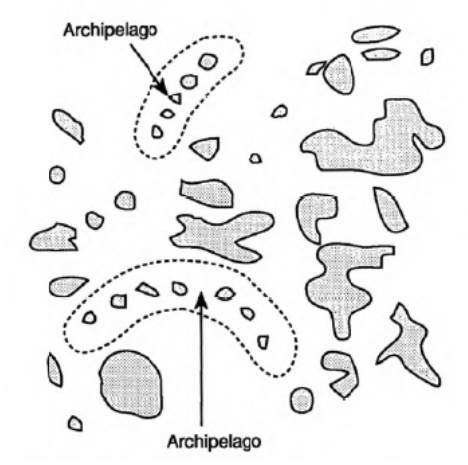


Figure 3: example of archipelagos that are deemed significant features for meso-scale pattern (taken from Müller and Wang 1992)

A specific case noted in Müller and Wang (1992) not solved by their methodology is the case of identifying archipelagos. Figure 3 demonstrates a case where an area-patch solution might overlook polygons that are at the micro (individual) level deemed too small to play a role in the area-patch problem, but analyzed at a meso-scale (group level) pattern are meaningful features to preserve.

The identification of archipelagos is considered to be one of the most difficult processes in automatic generalization (Bertin 1983, Kilpeläinen 2000, Steiniger and Weibel 2007). While several studies have demonstrated the ease for humans to identify clustered distributions of small groups of islands or lakes, there is not always agreement for the precise coincident extent of groups of polygon objects at certain scales (Bertin 1983, Steiniger et al. 2006). Various automated model generalization approaches for detecting clusters of polygons that could be considered archipelagos have been proposed (Müller and Wang 1992, Steiniger et al. 2006), but these papers only superficially explore the relationship between target generalization scale, aggregation and archipelagos. Specifically, there is a need to investigate the effects of including certain polygons to the class of meso objects called archipelagos and the relationship of that inclusion to subsequent aggregation steps and the resultant areal extent. In other words, if an algorithm chooses to include an island or lake structure in an archipelago, how will this affect the generalization outcome -as opposed to common approaches that would ignore the object?

1.4 Fuzzy Set Theory and the Concept of Archipelago

This paper presents a methodology to incorporate archipelago identification into generalization processes based on fuzzy set theory. Archipelagos are inherently vague geographic objects which are subject to Sorites paradox (Fisher 2000). A fuzzy set approach to geographic objects allows for the incorporation of

uncertainty in spatial relations and semantic meaning and would allow for a better understanding of the consequences for area-patch generalization (Winter 2000).

The concept of archipelagos is vague and therefore what constitutes an archipelago is open to debate. Concomitantly, how one defines the elements of an archipelago will affect the results of a generalized area-patch database. Since generalization is aimed at preserving the crucial characteristics of a map given the constraints of the output map environment, it is assumed that when aggregating, an algorithm needs to reflect the texture of the archipelago distribution. In order to properly aggregate the polygons significant elements of the pattern distribution need to be identified. Since an archipelago can be constituted of a range of polygons with different areal extents, it is assumed that certain polygons are more significant than others based on their size, aggregate numbers, and mean distance to each other. Such characteristics can be used to define fuzzy objects based on degrees of membership.

As an example Figure 4 depicts texture as a function of three elements, which could inherently describe the concept of an archetypical archipelago. The number of elements could be increased as needed. As mentioned previously, since it is assumed that archipelagos are vague geographic entities here the conceptual model is not understood to be strict. Generic "phenomenological" area-patch solutions are influenced by the intended type of map output, i.e. a walking map or a road map (Ormsby and Mackaness 1999).

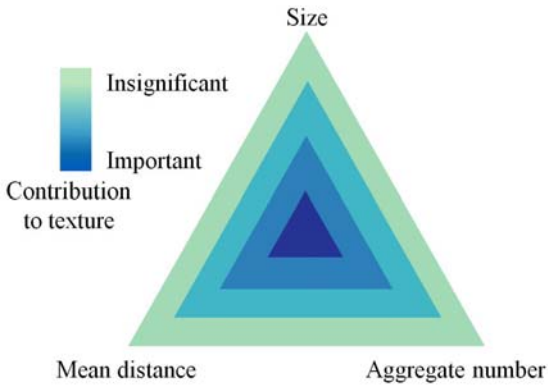


Figure 4: Univariate color gradient to illustrate the relationship between size, distance, and aggregate number of polygons for capturing archipelago texture.

2. DATA AND METHODS

2.1 Data

This research implements a fuzzy logic approach to archipelago identification based on both Müller and Wang (1992) and Ormsby and Mackaness (1999) to a hydrologic dataset from the National Hydrography Dataset (NHD). The dataset chosen for this research is a vector dataset of a subbasin (NHD # 03110201) of the Upper Suwannee River, which flows across the border between the states of Georgia and Florida. In this specific watershed the Suwannee River flows through a subbasin containing thousands of swamp/marshes and thousands of lake/ponds known as the Okefenokee Swamp.

The United States Geologic Survey (USGS) maintains and coordinates the NHD, a database of all the surface water of the United States. There are several versions of this database: high, medium, and local resolution (http://nhd.usgs.gov/data.html). The high-resolution dataset is compiled at a scale of 1:24,000 and covers all of the contiguous U.S. The medium-resolution is compiled semi-independently from the high-resolution dataset at 1:100,000. The local resolution is available only in select areas and is compiled at 1:4,800. All resolutions of the NHD constitute the surface water database of The National Map (TNM) and are designed to be a comprehensive dataset for GIS purposes (http://nhd.usgs.gov/).

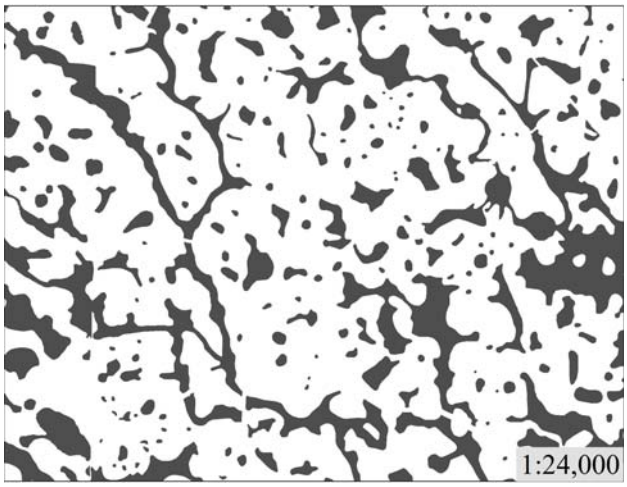


Figure 5 NHD high-resolution swamp/marsh data shown at the original compilation scale

Since the high and medium resolution NHD datasets are the only datasets that comprehensively catalogue the entire U.S. the high-resolution NHD was chosen as the dataset for this paper. Since the medium-resolution dataset is 1:100,000 the scale change area-patch solution proposed in this study is targeted for 1:100,000 to facilitate comparison of results in future work. As demonstrated in Figure 6 if you take the NHD high-resolution dataset of Figure 5 and display it at the target scale of 1:100,000, this results in a much too busy visual map display.

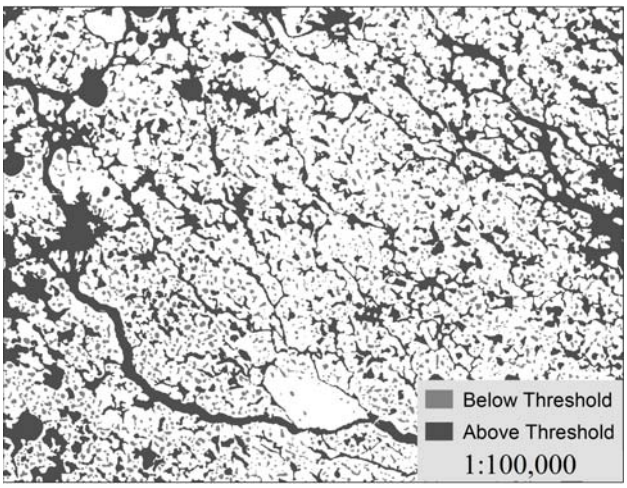


Figure 6: NHD high-resolution swamp/marsh data displayed at the medium resolution target scale.

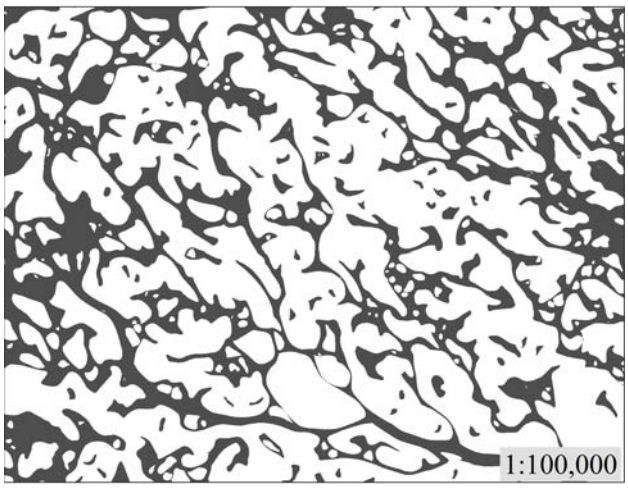


Figure 7: /Marsh area-patch solution utilizing a crisp delineation of feature areal contribution to pattern distribution.

2.2 Methodological Constraints

The methodology of this research is constrained in different aspects. First, this study focuses on NHD as a common dataset since it opens up the possibility for further work to be implemented on other area-patch problems within the NHD in order to develop common user GIS tools. Second, this research focuses exclusively on area-patch model generalization (Brassel and Weibel 1998). Therefore geometric constraints are taken from common USGS map production specs with the assumption that these generic metrics could be modified based on different map purposes.

The hydrography feature chosen for this study is the swamp/marsh waterbody class as delineated by the USGS for the NHD. The swamp/marsh waterbody underlies specific constraints for cartographic best practices for generalization such as a minimum size of 0.01 in^2 at map scale, which translates to a minimum feature size of ~64,500m^2 at the target scale of 1:100,000.

Figure 7 shows an example of a generalization of the polygons in Figure 6 after features that are too small for the target scale have been removed. Results are aggregated and then smoothed using the Polynomial with Approximation Exponential Kernel (PAEK) smoothing algorithm from ArcINFO 9.3. While this solution looks appropriate, visually, there is no accounting of archipelagos. Thus features whose size is below the defined constraint but collectively could constitute significant features are not included.

2.3 Generalization Procedures

As a first step in archipelago identification prospective centers of clustered swamp marshes that fall below the USGS size threshold of 64,500m^2 are identified using the Gi* hotspot analysis in ArcINFO 9.3 similar to Ormsby and Mackaness (1999). The result of the first pass of the Gi* tool is a Z score per polygon denoting the strength of spatial autocorrelation. Polygons, which were identified as cores of significant clusters ($p < 0.05$), were selected as candidate archipelago centers. Originally, the output of the Gi* tool did not give enough core areas and the remaining polygons were input to a second run of the Gi* tool in order to also include cluster centres of second order. Collectively these two-tiered core areas are named Seeds with the goal to catch the distribution of possible archipelagos. Upon visual inspection the output of these two runs of the G* tool was acceptable.

In order to identify the non-core polygons that participate in a perceived archipelago cluster, i.e. are proximate to the Seeds we carry out a second step. A table is generated in which for each Seed the closest objects within a 600 meter radius and their distances are recorded based on the NEAR function. These distances represent important elements in defining the prototype concept of a complete Archipelago. The search radius was chosen after several iterations and is based on the overall mean distance captured (e.g. a radius of 1000m only captures 13% more mean distances than 600m).

A third step is the derivation of the number of neighbours per Seed within the same search radius of 600m which is deemed to be the maximum distance accepted for an archipelago composition. Again, using the NEAR tool the total number of polygons per Seed is derived and recorded in a table.

2.4 Fuzzy Model and Prototype Concept

The basic idea of using fuzzy set theory in this study stems from the inherent vagueness of the phenomenon “archipelago”,

which lacks any definition in a semantic or quantitative sense. The core idea of this approach is to use semantic components that are known to be important for defining an archipelago and to quantify them by assigning fuzzy membership functions. This is referred to as Semantic Import (SI) (Burrough and McDonnell, 1998). The interplay of these relevant semantic concepts, which is operationalized by a fuzzy overlay, describes to what degree the prototypical concept of an archipelago is fulfilled expressed by a membership degree. Thus the "perfect" phenomenon of interest has to be defined using the same semantic components. Figure 4 shows this concept based on three semantic components that together define the vague object of interest.

These three semantic components or variables that together define a membership of a considered Seed to the concept of an archipelago: Areal extent of the Seed, mean distance between the Seed and its neighbours within 600 m and the number of neighbours also within 600m. It should be reminded that the Seeds from the G* analysis are thought to be core areas of archipelagos that once identified are used to capture the other polygons within the neighbourhood. Thus here it is attempted to find out which of the remaining polygons belong to the "anchored" archipelago.

For each attribute a membership function is assigned which allows association of the attribute to a fuzzy membership value resulting in a fuzzy set. These three fuzzy sets are then overlayed by using fuzzy logical operators. The result is a fuzzy set that defines the degree of membership to the concept archipelago. Different logic operators have been developed (Yager, 1980; Zadeh, 1965) to overlay two fuzzy sets. The optimal operator reflects best the behaviour of the system under consideration and the relationship between the semantic components (variables).

2.5 Fuzzy Membership Functions and Fuzzy Overlay

We assigned membership functions that are mathematically as simple as possible. Because archipelagos are vague and ill-defined objects we concentrated on the conceptual approach in this first attempt to model them and kept the analysis simple. Membership functions were chosen based on the results of the archipelago cluster analysis and above described constraints.

Our first semantic component assumes that within an archipelago larger features have a more significant contribution than smaller features. The membership function describes a simple linear relationship over the range of area between ~24,000m^2 and 64,500m^2.

The second semantic component takes into account proximity to other polygons. The closer the immediate neighbours are to the Seed the more significant is the contribution of the considered polygon to an archipelago. The left-trapezoidal membership function of this distance measure assigns membership values over the range 50-250m as a linear function with a slope (Figure 8 middle). For distances beyond the 50m threshold (minimum width delineation of 0.01 at map scale) features are assumed to merge and a membership value of 1 is assigned.

In order to implement a simple density measure as a third semantic component we count the number of polygons within a distance of 600m. The more polygons are found in this environment the higher the possibility that this polygon is part of an archipelago. Figure 8 bottom shows the corresponding membership function where 13 polygons and above are considered to be full members of the semantic component.

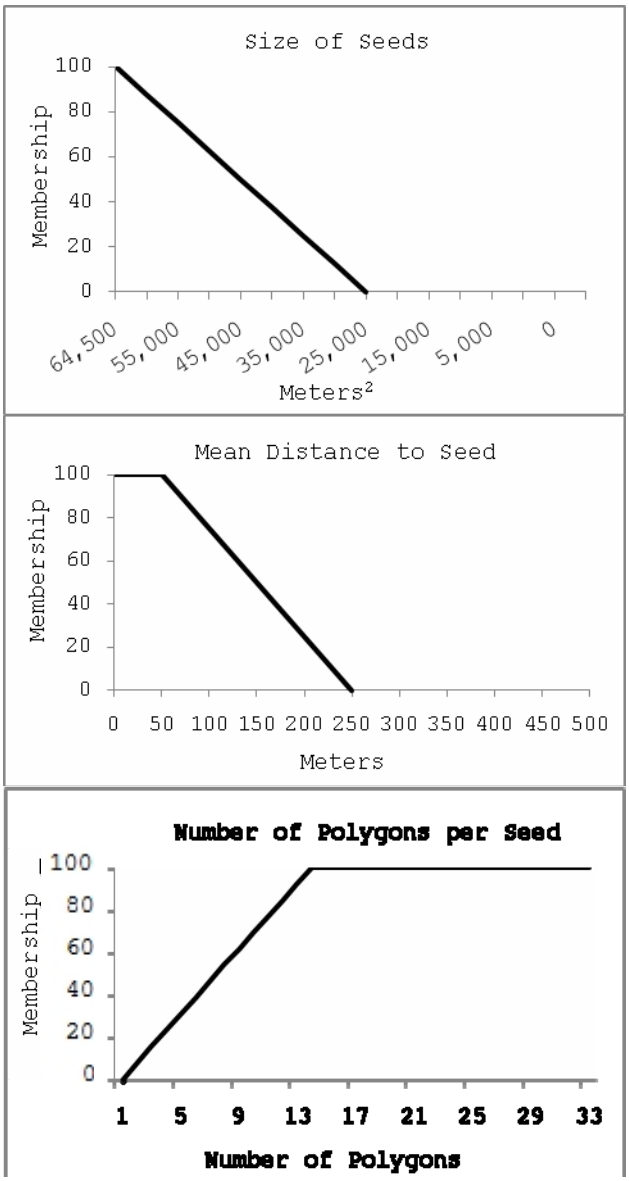


Figure 8: Semantic Import memberships per Seed.

We carried out fuzzy overlay by calculating the algebraic product between the fuzzy sets. The reason was that we needed a simple compensatory fuzzy operator, which weights the individual semantic components equally allowing each component to take effect but has also some conservative characteristics of an intersection operator. Fuzzy operators commonly used in GIScience are summarized in Robinson (2003).

As a result of this analysis each individual polygon obtains a membership value that expresses its degree of membership to

the prototype concept of archipelago. Using different alpha-cuts we tested the optimal threshold value above which the polygons are classified as members of an archipelago and implemented this in the final generalization process.

3. RESULTS

The final generalization is shown in Figure 9 where Seed polygons with acceptable membership percentages based on an alpha-cut of >20% given Figure 10 were accepted as archipelago members. Along with the Seeds, smaller polygons within 600m of the accepted Seed are aggregated as well. This therefore allows for a full range of polygons to be included in archipelago aggregates.

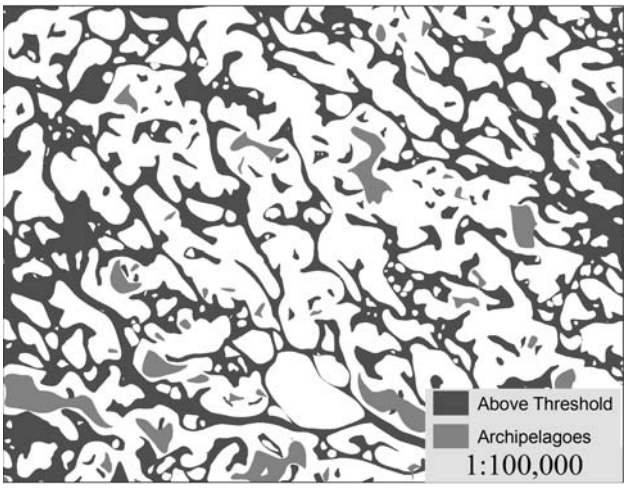
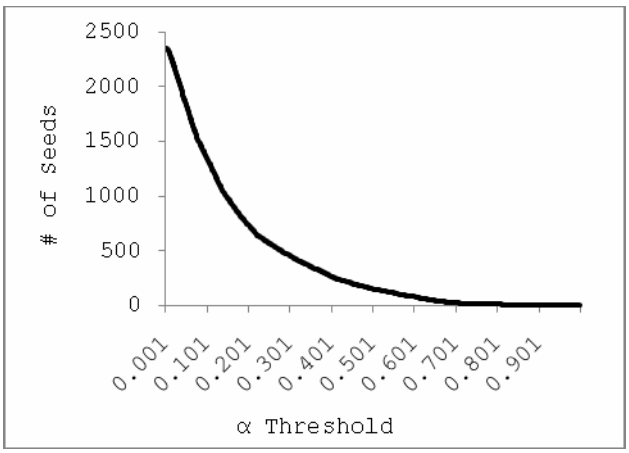


Figure 9: : Results of the aggregated and generalized significant archipelagos alongside the original generalized solution. The results were aggregated together and then smoothed using the PAEK smoothing algorithm set to a tolerance of 2000m.

The significant clusters, which resulted from the Gi* analysis and were used as core areas of archipelago candidates, can be seen in Figure 9. The visual inspection of this distribution indicates an acceptable amount of swamp/marsh polygons that provide a good basis for incorporation of archipelago-like compositions in the generalization. Figure 10 shows in detail the number of polygons included in the final generalization plotted as a function of possible alpha cuts used to accept the memberships as archipelago element. at an increment of 0.1%.



.

Figure 10: Number of polygons selected as archipelagos given a range of alpha cuts.

3.1 Discussion

This research is considered to improve the identification of archipelagos in a generalization process and thus to put forward Müller and Wang's (1992) results. However, there is still much more work to be done to improve this methodology. One major shortcoming of this research is the lack of use of different fuzzy overlay operators. How one combines semantic imports could greatly alter results especially if priority is given to one semantic import over another. Another shortcoming, which might be easily fixed, is that several polygons in figure 9 have artefact holes in their center, which does not reflect the character of the original swamp/marsh polygons. Future research could also explore different much more drastic scale jumps and whether it is easier to generalize on the generalized data or if it is better to generalize all scales from the highest resolution data. Ultimately this research is directly applicable to multi-scale generalization approaches to area-patch problems especially if one considers the possibility of only needing to modify an alpha cut or a semantic import level to generalize to a different scale.

4. CONCLUSION

In conclusion, there are still many options left to explore in relation to this fuzzy logic approach to archipelagos and the area-patch generalization of the Upper Suwannee subbasin. However, the fact that this methodology was implemented using widely available out of the box GIS software and data processing tools gives credence to the usefulness of this approach. This methodology allows for the identification of vague objects such as archipelagos while offering flexibility necessary for a wide variety of NHD users who might need to generalize their data.

4.1 References and/or Selected Bibliography

Ahlqvist, O., Keukelaar, J., and Oukbir, K., 2003. Rough and fuzzy geographical data integration. *International Journal of Geographical Information Science*, 17(3), pp. 223-234.

Beard, K., 1991. Constraints on rule formation. In: *Map generalization : making rules for knowledge representation*. B. P. Buttenfield and R. B. McMaster, Longman.

Bertin, J., 1983. *Semiology of graphics*. University of Wisconsin Press, Madison, pp. 304-307.

Brassel, K., and R., Weibel 1988. A review and conceptual framework of automated map generalization. *International Journal of Geographical Information Science*, 2(3), pp. 229-244.

Bundy, G., Jones, C., and Furse, E., 1995. Holistic Generalization of Large-Scale Cartographic Data. In: *GIS and generalization : methodology and practice*. J. C. Müller, J. P. Lagrange and R. Weibel. London ; Bristol, PA, Taylor & Francis, pp. 106-119.

Burrough, P., and McDonnell R., 1998. *Principals of Geographical Information Systems*. Oxford University Press.

Buttenfield, B.P., and D., Mark, 1991. Expert Systems in Cartographic Design. In: *Geographic Information Systems: the Microcomputer and Modern Cartography*. D. Taylor; Pergamon Press, Oxford, pp.129-150.

Deng, Y., and Wilson, J. P., 2008. Multi-scale and multi-criteria mapping of mountain peaks as fuzzy entities. *International Journal of Geographical Information Science*, 22(2), pp. 205-218.

Fisher, P., 1992. First Experiments in Viewshed Uncertainty: Simulating Fuzzy Viewsheds. *Photogrammetric Engineering & Remote Sensing*, 58(3), pp. 345-352.

Fisher, P., 2000. Sorites paradox and Vague Geographies. *Fuzzy Sets and Systems*, 113(1), pp. 7-18.

Kilpeläinen, T., 2000. Knowledge Acquisition for Generalization Rules. *Cartography and Geographic Information Science*, 27(1), pp. 41-50.

Müller, J. C., and Z. S., Wang, 1992. Area-patch generalization: a competitive approach. *Cartographic Journal*,*s* 29(2), pp. 137-144.

Müller, J. C., R., Weibel, et al., 1995. Generalization: state of the art and issues. *GIS and generalization : methodology and practice*,. J. C. Muller, J. P. Lagrange and R. Weibel, London ; Bristol, PA, Taylor & Francis, pp. 3-17.

Ormsby, D., and W., Mackaness, 1999. The Development of Phenomenological Generalization Within an Object-oriented Paradigm. *Cartography and Geographic Information Science*, 26, pp. 70-80.

Robinson, V., 2003. A Perspective on the Fundamentals of Fuzzy Sets and their Use in Geographic Information Systems. *Transactions in GIS*, 7(1), pp. 3-30.

Ruas, A., 2000. The roles of meso objects for generalisation. *Proceedings of the International Symposium on Spatial Data Handling*, Beijing, China, sec 3b.

Steiniger, S., et al., 2006. Recognition of island structures for map generalization. *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, Arlington, Virginia, ACM-GIS'06.

Steiniger, S., and R., Weibel, 2007. Relations among Map Objects in Cartographic Generalization. *Cartography and Geographic Information Science*, 34(3), pp. 175-197.

Töpfer, F., and W., Pillewizer, 1966. The Principles of Selection. *The Cartographic Journal*, 3, pp. 10-16.

Varzi, A., 2001. Vagueness in geography. *Philosophy and Geography*, 4, pp. 49–65.

Winter, S., 2000. Uncertain topological relations between imprecise regions. *International Journal of Geographic Information Science*, 14(5), pp.411-430.

Yager, R., 1980. On a general class of fuzzy connectives. *Fuzzy Sets and System*,. 4(3), pp. 235-242.

Zadeh, L.A., 1965. Fuzzy sets. *Information and Control*, 8, p. 338-353.