An AMOEBA procedure for visualizing clusters

Marta Jankowska[1], Jared Aldstadt[2], Arthur Getis[1], John Weeks[1], Grant Fraley[1]

[1]San Diego State University/ Department of Geography
mjankows@rohan.sdsu.edu

[2]University at Buffalo/ Department of Geography
geojared@buffalo.edu

INTRODUCTION

The AMOEBA procedure, or A Multidirectional Optimum Ecotope-Based Algorithm, as developed by Aldstadt and Getis (2006) is designed to identify hot and cold spots in mapped data by assessing the spatial association of a mapped unit to surrounding units. Because of its emphasis on statistically significant hot and cold spots, AMOEBA was not designed to exhaustively assign all units to clusters. In the work discussed here, we develop an AMOEBA procedure that is able to assign all units to spatial clusters of similar characteristics and create useful visualizations of the results. This version of AMOEBA includes additions to the original algorithm that are designed to define and control cluster size, and is relevant for a variety of applications where visualizations of reasonably homogenous sub-regions are required. We demonstrate our procedure using data from the 2000 Ghana Census to identify social neighborhoods in Accra, Ghana.

VISUALIZING AMOEBA

AMOEBA is able to map clusters of high and low values by creating a spatial weights matrix based on the Getis-Ord *Gi** (Ord and Getis, 1995), or any other local statistic. The algorithm develops a cluster from a selected seed location by evaluating *Gi** values of all locations surrounding the seed. If the neighbor increases the *Gi** value, it is included into the forming cluster; if it does not increase the *Gi** value, it is excluded. By placing a cluster seed in every location, AMOEBA becomes exhaustive. Yet, if the purpose is to assign each unit to a cluster, not only to hot and cold spots, this results in significant cluster overlap as each seed is allowed to grow into locations that belong to earlier clusters.

With the overlapping of clusters, new avenues for visualization of clusters must be applied. This version of AMOEBA outputs a number of statistics for each observation including the *Gi*max* value, which is the largest *Gi** value of all the clusters that include the observation. When

mapped, the *Gi*max* statistic provides a picture of contrast in spatial autocorrelation, and by visually merging adjacent clusters with similar *Gi** values it provides boundaries for clusters. An example of mapped AMOEBA *Gi*max* values is displayed in Figure 1. This figure is derived from running AMOEBA on the first principle component of four variables; three socioeconomic (SES) variables, and one slum variable. These variables are taken from the 2000 Ghana Census for 1,717 enumeration areas (EAs) in Accra, Ghana. The resulting *Gi*max* values displayed in Figure 1 are divided into five value ranges, with high absolute values of *Gi*max* indicating high spatial autocorrelation. In this figure, negative values indicate spatial autocorrelation of individuals working in the professional workforce, and positive values indicate spatial autocorrelation of individuals with low literacy, high rates of working in the informal sector, and high percentages of people living in slum like housing.

Visualization of clusters is highly controlled by the divisions of *Gi*max* values. Clusters in Figure 1 are relatively large, and might be broken down if *Gi*max* is divided into more value ranges. But areas with extreme autocorrelation will have the same, or very similar, *Gi*max* values precluding the possibility of breaking these clusters up in the visualization process. For example, in Figure 1 all observations in the white cluster running down the center of Accra have a *Gi*max* value of -19.64. This cluster boundary is defined by *Gi*max*, and cannot be manipulated.

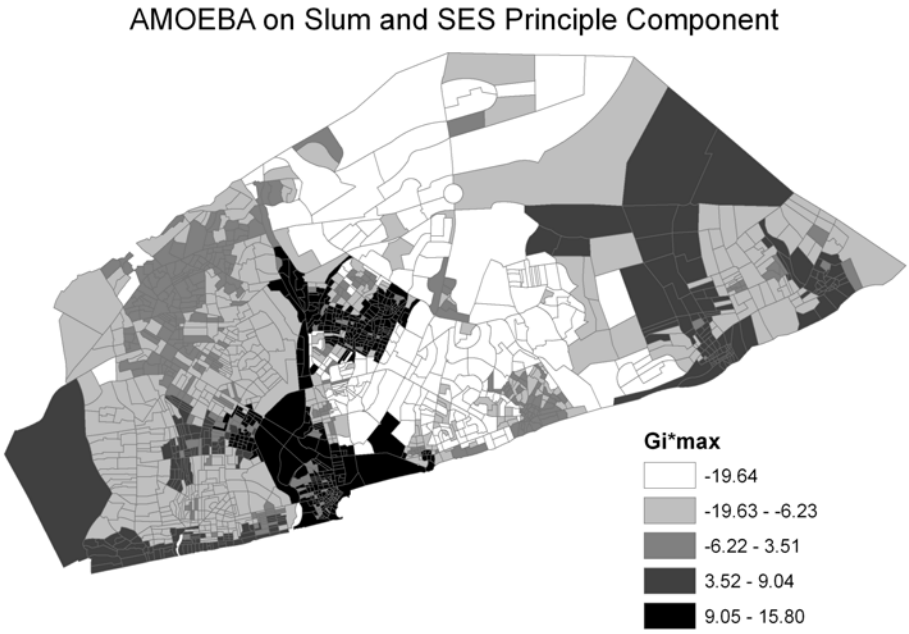

Fig. 1: AMOEBA *Gi*max* of first principle component of SES and slum.

CONTROLLING CLUSTER SIZE

As demonstrated in Figure 1, in cases of extreme autocorrelation, *Gi*max* is able to delineate an exact cluster boundary. But there may be circumstances where more flexibility in visualization is desired. For example, if the goal is to create spatially agglomerated areas from EAs, such as neighborhoods, some areas in Figure 1 would be too large to be realistic neighborhoods. In this case user control of cluster size is desirable. Kulldorff et al. (1998) limit cluster size by population, while Tango and Takahashi's (2005) allows the analyst to limit number of units in a cluster. Two similar features are introduced in AMOEBA: the first gives control over the maximum number of observations in a cluster. The second allows the analyst to place a threshold on a variable of choice such as area or population, where the cluster cannot exceed the sum of the selected variable. Figure 2 demonstrates AMOEBA *Gi*max* values with an observation limit of 25 EAs, and an area threshold of the sum of the two largest EAs in Accra. It is immediately noticeable that the white cluster down the center of Accra from Figure 1 is broken down into many smaller clusters, creating more feasible neighborhood sizes.

With these new additions to the original AMOEBA procedure, all observations are included into clusters allowing visualization of autocorrelation over an entire area. By introducing limitations on cluster growth, further flexibility in cluster visualization is achieved as more options are available to the user.

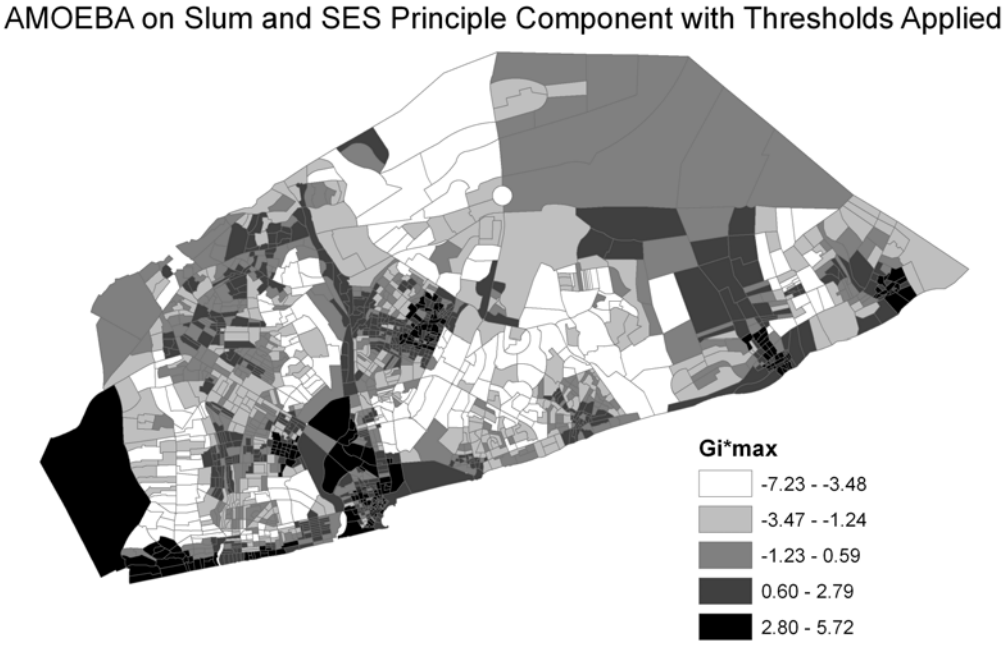

Fig. 2: AMOEBA *Gi*max* values for SES and slum; EA and area thresholds introduced.

ACKNOWLEDGMENTS

This research was supported by Grant Number R01HD054906 from the National Institute of Child Health and Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Child Health and Human Development or the National Institutes of Health.

REFERENCES

Aldstadt, J. and A. Getis (2006). "Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters." Geographical Analysis 38: 327-343.

Kulldorff, M., K. Rand, et al. (1998). SaTScan v 2.1: Software for the Spatial and Space-Time Scan Statistics. Bethesda, MD: National Cancer Institute.

Ord, J.K. and A. Getis (1995). "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application." Geographical Analysis 27: 286-306.

Tango, T. and K. Takahashi (2005). "A flexibly shaped spatial scan statistic for detecting clusters." International Journal of Health Geographics 4(11).