Working Paper 2009-01

Spatially Lagged Choropleth Display

Alan T. Murray

SPATIALLY LAGGED CHOROPLETH DISPLAY

Alan T. Murray

School of Geographical Sciences and Urban Planning
Arizona State University
Tempe, AZ 85287
USA
Email: atmurray@asu.edu

ABSTRACT
Choropleth display of spatial information is a fundamental feature of mapping and geographic information system technologies. There has long been a desire to impart some spatial influence in the class selection and delineation process of choropleth display. This paper presents an approach for representing the spatial influence of neighboring areas in the creation of choropleth classes. The usefulness of this approach is explored using suburb level crime statistics for Brisbane, Australia.

1. INTRODUCTION
The study of any phenomenon across space invariably begins with a choropleth display of the attribute(s) of interest. As an example, if we are interested in the distribution of crime in a city like Brisbane, Australia, we would likely start by looking at the crime levels in suburbs of this city. A standard way to evaluate the spatial variation of crime statistics using a geographic information system (GIS) or mapping software would be to color each suburb based upon its crime rate in relation to other suburbs in the region. Figure 1 illustrates the distribution of property crime per 1000 residents for 1996 in each suburb in the city of Brisbane. The legend gives some indication of the range of crime rates in this region. Figure 1 is known as a choropleth display.

Substantial research has been devoted to choropleth display over the past 50 years (Jenks 1963; Evans 1977; Coulson 1987; Dent 1990). The focus of this research has been to identify effective methods for depicting differences in displayed attributes. Developed choropleth methods take the range of attribute values, say total population in each suburb as an example, and use a significantly smaller number of groups or classes (typically between 4-7 as suggested in Dent 1990) to depict regional variation. Although spatial patterns and relationships may be inferred in the distribution observed in Figure 1, the classification process is aspatial. That is, classification groups (or break points) are determined by the distribution of the attribute being displayed, irrespective of their relative spatial location.

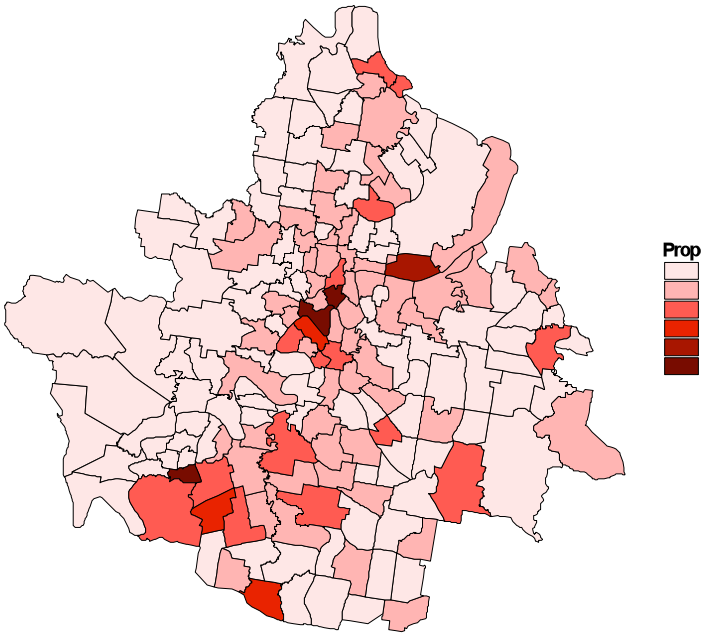

Figure 1. Choropleth display of property crime rates.

There has long been interest in extending the basic attribute classification process to reflect spatial relationships between display units (Jenks and Caspall 1971). As an example, we might find upon further investigation that two neighboring units belonging to different classes in Figure 1 may actually be quite similar in attribute value. Thus, given the relative proximity of the units with similar attribute values, in this case property crime rates, it may be reasonable to have such units in the same class, because there is essentially no attribute variation in this local area. The significance here is that units may be in different classes, which suggests a change in the regional distribution of property crime, but they may in fact be somewhat similar. The power of choropleth display is obviously substantial given the ability to influence the interpretation of regional change. An initial attempt to incorporate space into the choropleth display process was devised to account for boundary relationships in order to alter class composition (see Jenks and Caspall 1971; Monmonier 1972; Cromley 1996). This was a clever and interesting way to indirectly account for spatial relationships. There are, however, alternative approaches that may be worth pursuing. This is particularly true for cases where one is interested in local hot (or cold) spots. That is, both attribute similarity and spatial proximity have some significance in how units should be classified, displayed and interpreted.

The paper explores the use of an indirect approach for including spatial relationships in the creation of choropleth classes. Specifically, we propose the traditional form of spatial lag used in spatial statistics as a means for altering the interpretation of attribute values in class delineation. Two multi-objective clustering models are proposed for integrating attribute and spatial lag. These approaches may be considered extensions to the median based non-hierarchical classification problem. Comparative results are presented for the analysis of property crime in Brisbane, Australia.

2. SPATIAL LAG

In the analysis of spatial information it is often important, if not critical, to know what is happening around a particular location. The typical assumption of independence between observations of interest in classic statistical techniques has long been known to be problematic for spatially referenced objects (Griffith and Amrhein 1997). Specifically, the existence of spatial autocorrelation may in fact alter significance levels and reduce interpretation of results. One approach for dealing with spatial autocorrelation has been to utilize spatial lag in various statistical techniques such as multivariate linear regression. Spatial lag essentially represents an averaging of the attributes around a particular location. As an example, if we looked at one of the suburbs shown in Figure 1 (depicting property crime rates), the spatial lag for this suburb would be the average rate of property crime for all of the suburbs which are defined as neighboring this suburb. Typically, neighbors are those suburbs which share a common boundary with the suburb being considered. This is now formally specified.

Consider the following notation:

i = index of areas;
f_i = attribute measure;
l_i = spatial lag of attribute measure;
N_i = spatial neighbors of area *i*.

Areas correspond to suburbs in Figure 1 and the attribute measure is property crime. Neighbors are defined here as those suburbs sharing a common border or point, but alternative interpretations such within a specified distance could be utilized with any loss of generality. Using this notation, the spatial lag for an area *i* is as follows:

$$l_i = \frac{\sum_{j \in N_i} f_j}{|N_i|} \quad (1)$$

Thus, the spatial lag is nothing other than an average value of the neighbors to a particular area. It is worth noting that the spatial neighbors of an area do not include the area itself. The spatial lag of an area is a summary indicator of what is happening in a relative location without tracking the exact relationships between neighboring areas. As such, it is a proxy for the spatial similarity or difference of an area and its neighbors (to the extent that one defines units as being neighbors of each other).

3. SPACE AND ATTRIBUTE SIMILARITY

It is now possible to define the similarity of two areas in terms of their attribute measures as well in terms of their spatial lag. This similarity is denoted as:

s_{ij} = similarity of areas *i* and *j*.

One approach for integrating attribute and spatial lag differences between two areas is as follows:

$$s_{ij} = \left|\left(w_a f_i + w_d l_i\right) - \left(w_a f_j + w_d l_j\right)\right| \quad (2)$$

where

w_a = importance weight for attribute similarity;
w_d = importance weight for spatial lag.

This similarity measure utilizes weights to combine the attribute and spatial lag values for one area and compare it to another area. As we are interested in their relationship, we take the absolute values of their difference.

An alternative approach for integrating attribute and spatial lag similarity involves weighting the two components after they have been compared. This would utilize two measures of similarity:

a_{ij} = attribute similarity of areas i and j;
d_{ij} = spatial lag similarity of areas i and j.

These individual similarity measures may be formalized as follows:

$$a_{ij} = \left| f_i - f_j \right| \quad (3)$$

$$d_{ij} = \left| l_i - l_j \right| \quad (4)$$

Integrating attribute and spatial lag similarity would then involve the use of the above weights. This will be formally detailed in the following section. It is worth noting that the integrated similarity measure defined in (2) obviously differs from a weighted combination of the attribute and spatial lag similarity measures defined in (3) and (4) respectively. Incorporating these measures in an optimization based classification model for choropleth display will now be developed.

4. SPATIALLY LAGGED CLASSIFICATION

The most widely advocated approach for choropleth display is the natural breaks method, which minimizes the variance of attributes within classes. The classes depicted in Figure 1 were obtained using the natural breaks choropleth display option in ArcView (version 3.2). This approach is defined as being a single dimensional non-hierarchical classification/clustering problem and may be solved using the technique of Fisher (1958) to determine the appropriate classes for depicting an attribute in a choropleth display. A related classification approach is to utilize a median clustering model in order to minimize within class difference (Monmonier 1973; Cromley 1996; Murray and Estivill-Castro 1998; Murray and Grubesic 2002). The developed classification models for choropleth display in this paper are based upon the use of the median clustering approach. The first model incorporates spatial lag in the classification of attributes in order to reflect the spatial variation around areas.

Some additional notation will be utilized:

j = index of potential medians (same as i);
p = number of classes to be identified.

Decision variables:

$$x_j = \begin{cases} 1 \text{ if class median } j \text{ is selected} \\ 0 \text{ otherwise.} \end{cases}$$

$$z_{ij} = \begin{cases} 1 \text{ if area } i \text{ is in class } j \\ 0 \text{ otherwise.} \end{cases}$$

Using the above notation, it is possible to structure a class selection model with objectives that simultaneously maximize attribute and spatial lag homogeneity.

Spatially Lagged Median Classification (SLMC)

Minimize
$$Z = \sum_i \sum_j s_{ij} z_{ij} \tag{5}$$

Subject to:

$$\sum_j z_{ij} = 1 \quad \forall i \tag{6}$$

$$\sum_j x_j = p \tag{7}$$

$$z_{ij} \leq x_j \quad \forall i, j \tag{8}$$

$$\begin{aligned} z_{ij} &= (0,1) \quad \forall i, j \\ x_j &= (0,1) \quad \forall j \end{aligned} \tag{9}$$

The objective (5) of the SLMC is to minimize the total weighted within group difference of selected classes. The constraints of the SLMC are standard median model conditions which ensure that classes are structured in a meaningful way. Constraint (6) ensures that each area is included in a class. Constraints (7) and (8) require that only p classess be generated. Constraints (9) impose integer restrictions on decision variables. Other spatial clustering models, such as those discussed in Murray and Estivill-Castro (1998), could be readily adapted to include attribute and spatial lag as is done in the SLMC.

A feature of this classification model is that the importance weights, w_a and w_d, may be varied to alter s_{ij}. Doing this will result in pseudo-tradeoff solutions. More will be said about this point later in the paper. An alternative classification model may be structured to incorporate attribute and spatial lag similarity more distinctly.

Bicriterion Spatially Lagged Median Classification (BSLMC)

Minimize
$$Z = w_a \sum_i \sum_j a_{ij} z_{ij} + w_d \sum_i \sum_j d_{ij} z_{ij} \tag{10}$$

Subject to:

Constraints (6)-(9)

The only difference between the BSLMC and SLMC is the objective (10). In the BSLMC the objective is to minimize the total weighted attribute similarity and to minimize the total weighted spatial lag similarity in selected classes. The constraints for the BSLMC are the same as those for the SLMC. The objective of the BSLMC is explicitly multi-objective. Thus, it is possible to analyze tradeoffs associated with varying the weights w_a and w_d.

A major interest in this paper is to explore the extend to which both of these classification models is capable of representing spatial variation in the creation of choropleth display classes. In addition, there is also interest in identifying performance differences between these two approaches.

5. APPLICATION RESULTS

As stated early in this paper, rates of property crime per 1000 residents in the 178 suburbs of Brisbane, Australia for 1996 are investigated. Lagrangian relaxation with branch and bound (see Murray and Gerrard 1997) was utilized for solving the Spatially Lagged Median Classification (SLMC) and Bicriterion Spatially Lagged Median Clustering (BSLMC) problems optimally on a Pentium III/600 personal computer. These models were implemented as dynamic link libraries (DLL) compiled using Digital Visual Fortran (version 6.0) and integrated in ArcView (version 3.2) using Avenue scripts. Reported results for the SLMC and BSLMC problems in this paper are optimal to within 0.1% using Lagrangian relaxation. The time required to solve these problems was generally between 0-10 seconds, but there was an instance where 22 seconds was needed.

A global measure of the degree of spatial association between area attributes is the Moran's I statistic (see Griffith and Amrhein 1997). Using SpaceStat version 1.90, Moran's I for property crime was found to be 0.102 with a standard normal z-value of 2.37 ($p<0.02$). Given this, some amount of spatial similarity in crime rates does exist in Brisbane. This suggests that the use of the SLMC to influence class structure in order to better reflect spatial relationships is warranted. If a relatively low weight is given to spatial lag, then the attribute has a greater influence in class structure, but this may be altered somewhat depending upon attribute values in neighboring areas. An example of this is depicted in Figure 2, where the attribute weight is 1.0 and the weight for spatial lag is 0.1. Figure 2 shows some modification to the classes depicted in Figure 1. In addition to the suburbs noted in Figure 2, suburbs in the city center and in the far east have changed classes. The reason for this is that crime rates in these suburbs were relatively close to neighboring areas. Thus, the inclusion of spatial lag altered class delineation in order to account for rates in neighboring suburbs. The ability to alter weights in specifying similarity in the SLMC allows us to assess how class structure changes when spatial lag has greater influence. Figure 3 depicts the choropleth display for the SLMC using weights of 0.4 for the attribute and 1.0 for spatial lag. What may be inferred in Figure 2 is greater spatial contiguity of identified classes. In addition, a north-south corridor of relatively high property crime rates appears to be particularly pronounced in Figure 3. Figures 1-3 each appear to impart some insight in the patterns of property crime in Brisbane.

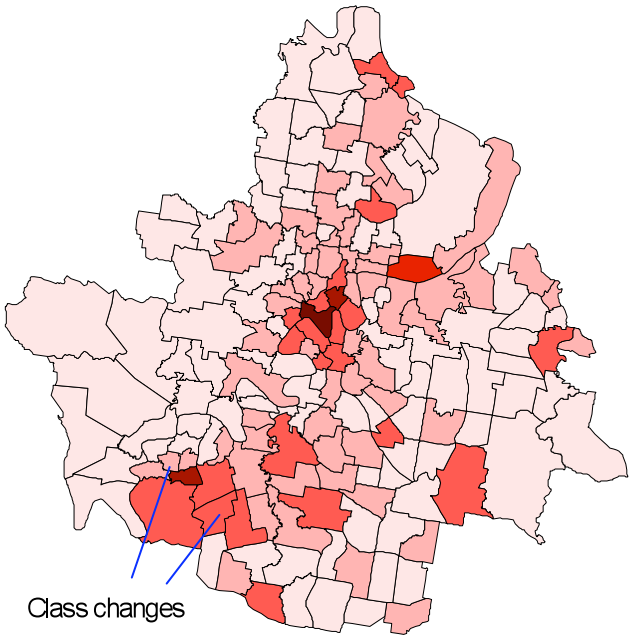

Figure 2. SLMC choropleth display of property crime rates (w_a=1.0 and w_d=0.1).

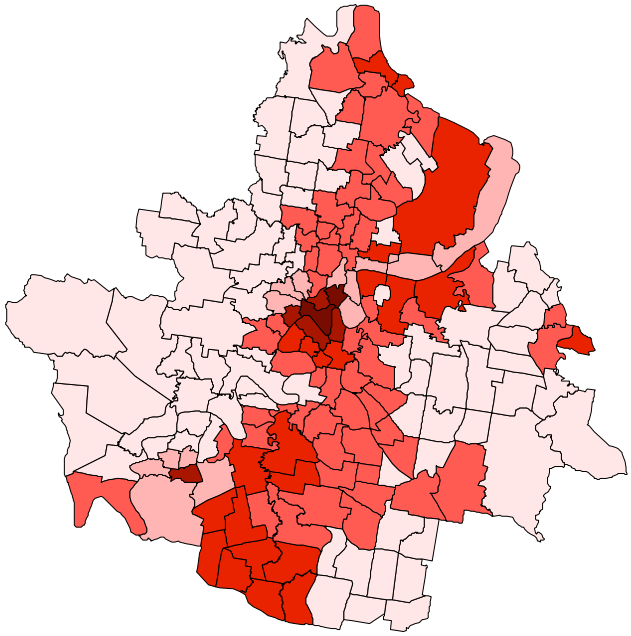

Figure 3. SLMC choropleth display of property crime rates (w_a=0.4 and w_d=1.0).

The BSLMC is also structured to influence class creation in choropleth display by incorporating attribute and spatial lag for determining similarity between areas. As the BSLMC is explicitly multi-objective, tradeoff solutions may be identified and evaluated. One such tradeoff solution is shown in Figure 4 for a relatively low spatial lag weight. Also displayed in Figure 4 is the non-inferior tradeoff curve associated with varying the attribute and spatial lag weights. Thus, along the x-axis is the spatial lag contribution to the BSLMC objective function and on the y-axis is the attribute contribution to the objective function (property crime in this case). The highlighted tradeoff solution corresponds to the choropleth display depicted in Figure 4 using the indicated weights. One noteworthy change in suburb classes is Geebung (highlighted in Figure 4). The property crime rate is 5.71 in Geebung and neighboring suburbs have somewhat similar rates (6.61 for Boondall; 8.44 for Virginia; 8.46 for Chermside; and 9.69 for Zillmere). In Figure 1 these suburbs are in a different class than Geebung. Using the BSLMC, we find that there is in fact some basis to suggest that Geebung should actually belong to the same class as these other suburbs. Similar class changes may also be found in Figure 4. What is particularly interesting in this classification is that it is somewhat different to that found using the SLMC in Figure 2. To further explore this, we can evaluate an instance where the weight for spatial lag is greater than that for property crime. Another tradeoff solution is shown in Figure 5 for an attribute weight of 0.4 and a spatial lag weight of 1.0. The inserted tradeoff curve identifies the relationship of this solution to other solutions (the one shown in Figure 4 in particular). As with Figure 3, we also see more spatial clustering in Figure 5 (as compared to Figures 1 or 4), but the patterns are different between Figures 3 and 5. Thus, although the resulting patterns produced by the SLMC and the BSLMC appear to be related, they are in fact different.

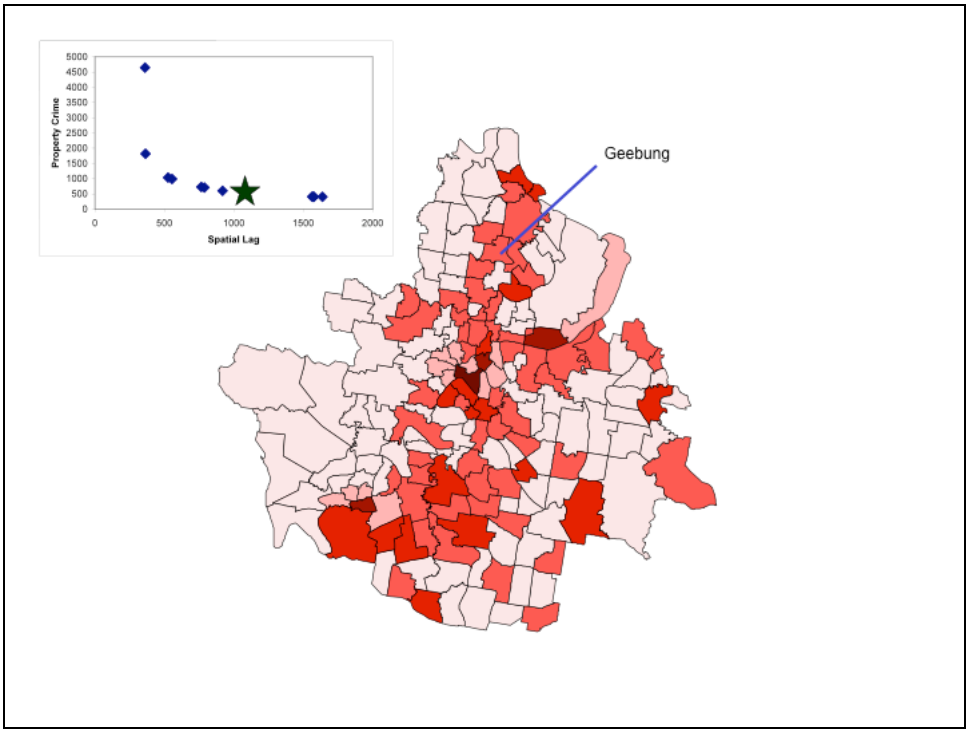

Figure 4. BSLMC choropleth display of property crime rates (w_a=1.0 and w_d=0.3).

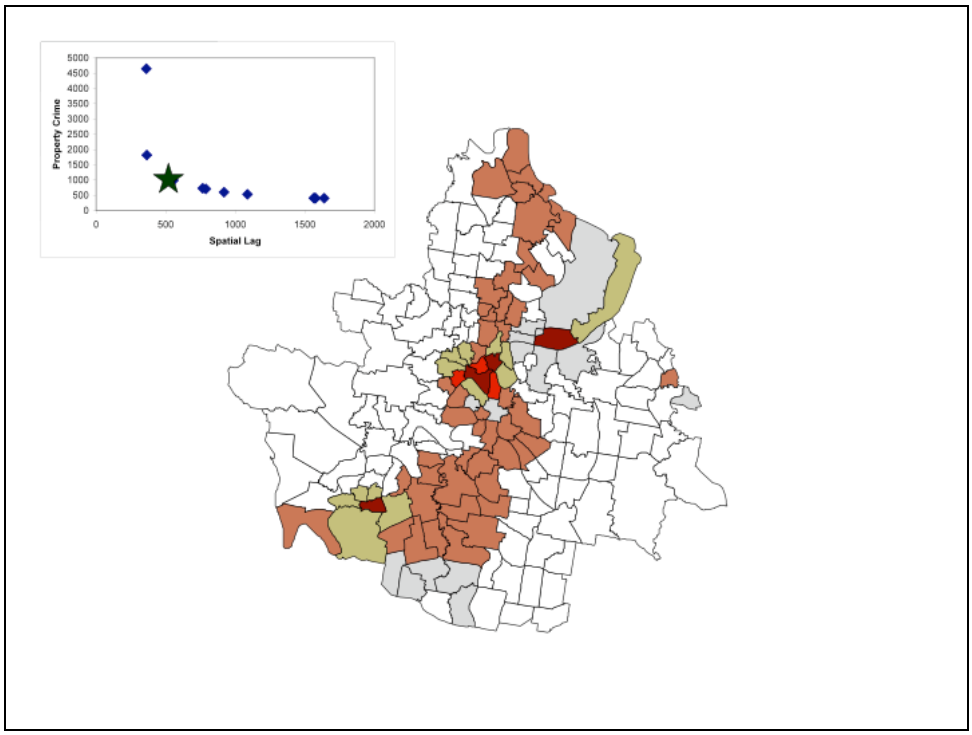

Figure 5. BSLMC choropleth display of property crime rates (w_a=0.4 and w_d=1.0).

6. DISCUSSION

The results of the SLMC and the BSLMC are both interesting and informative. Further analysis of each display highlights how property crime is distributed in Brisbane. From a planning and management perspective, the ultimate goal is to be able to make inferences about what is taking place and why. Are there geographic features which somehow correlate to either high or low property crime rates? Are there hot spots of crime activity? Identifying significant spatial clusters is part of being able to address these questions. Two relatively clear observations to be inferred from Figures 1-5 is the low rates of crime in the western and south-eastern suburbs as well as the high rates of property crime in and around the city center.

The inclusion of spatial lag in choropleth classification provides the ability to alter class groups based upon attribute values in neighboring areas. This is an appealing feature of both the SLMC and the BSLMC models. A valuable property of the spatial lag measure is that totally contiguous and spatially compact class groupings are never produced, even when spatial lag receives a dominate weighting. This would not be the case if a distance proximity measure was used in place of spatial lag. Thus, spatial lag appears to lessen the impact and influence of space in class creation.

A point raised earlier in the paper was that the SLMC does not produce non-inferior tradeoff solutions. This may be seen in Figure 6, which shows the attribute and spatial lag contributions to the objective function. These were obtained by varying attribute and spatial lag weights. In contrast to the tradeoff curves depicted in Figures 4 and 5, Figure 6 illustrates that varying the weights in equation (2) does not result in non-dominated solutions. This may be explained by the non-linear form of equation (2), since an absolute value of the differences is utilized in similarity specification. The SLMC is probably the most representative of how spatial lag would be included in the traditional 1-dimensional classification process. Thus, it is important to recognize that it actually produces dominated

tradeoff solutions. That is, group classifications exist which have an equivalent average property crime rate (see Figure 6), as an example, but these classifications have varying average spatial lag. What we are interested in is only the minimum average spatial lag classification as this would represent the non-inferior tradeoff solution. Given that this happens, this may in fact be an unappealing feature of the SLMC. As a result, the BSLMC would likely represent a better choropleth display alternative.

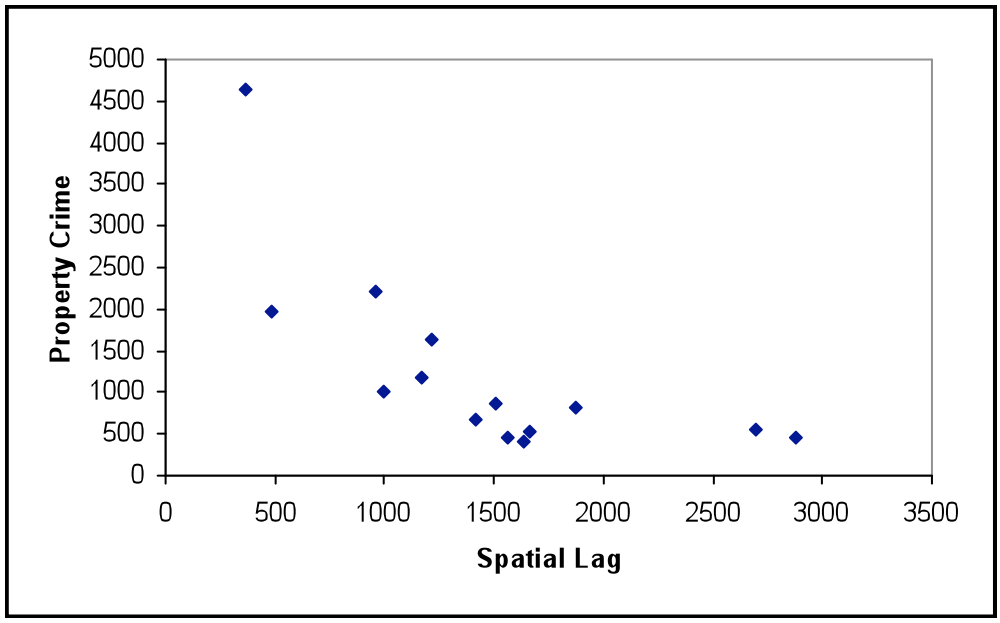

Figure 6. Tradeoff solutions associated with the SLMC.

7. CONCLUSIONS AND FUTURE DIRECTIONS

This paper has developed two alternative approaches for indirectly representing spatial relationships in choropleth class creation. This has been done using a median based clustering model which incorporates attribute and spatial lag similarities. To a certain degree, the inclusion of spatial lag may be viewed as a type of spatial filtering process for identifying group classifications. Previous research in choropleth display has noted the need for representing spatial significance in the classification process. Clearly this has been accomplished using spatial lag in the developed classification models. The application results illustrated the ability of both the SLMC and the BSLMC to alter class structure in a tempered way based upon the attribute values of neighboring areas. This is quite valuable for assessing the regional variation of crime rates as well as most other area attributes.

A number of areas for future research may be identified based upon this research. There is clearly a need for more understanding of the relationship between the use of spatial lag versus the boundary type approach detailed in Jenks and Caspall (1971) and more recently in Cromley (1996). Further, similar comparative work is needed for the use of explicit spatial relationship measures such as distance between areas in choropleth classification. A final area for future research is to explore the relationship between these classification techniques and spatial statistical approaches like local indicators of spatial association (see Anselin and Bao 1997).

REFERENCES

Anselin, L. and Bao, S. (1997). "Exploratory spatial data analysis linking SpaceStat and ArcView." In *Recent Developments in Spatial Analysis*, edited by M. Fischer and A. Getis, 35-59 (Berlin: Springer-Verlag).

Coulson, M. (1987). "In the matter of class intervals for choropleth maps: with particular reference to the work of George Jenks." *Cartographica* **24**, 16-39.

Cromley, R. (1996). "A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data." *International Journal of Geographical Information Systems* **10**, 405-424.

Dent, B. (1990). *Cartography: Thematic Map Design*, second edition (Dubuque: Wm. C. Brown Publishers).

Evans, I. (1977). "The selection of class intervals." *Transactions of the Institute of British Geographers* **2**, 98-124.

Fisher, W. (1958). "On grouping for maximum homogeneity." *Journal of the American Statistical Association* **53**, 789-798.

Griffith, D. and Amrhein, C. (1997). *Multivariate Statistical Analysis for Geographers* (New Jersey: Prentice-Hall).

Jenks, G. (1963). "Generalization in statistical mapping." *Annals of the Association of American Geographers* **53**, 15-26.

Jenks, G. and Caspall, F. (1971). "Error on choroplethic maps: definition, measurement, reduction." *Annals of the Association of American Geographers* **61**, 217-244.

Monmonier, M. (1972). "Contiguity-biased class-interval selection: a method for simplifying patterns on statistical maps." *Geographical Review* **62**, 203-228.

Monmonier, M. (1973). "Analogs between class-interval selection and location allocation models." *The Canadian Cartographer* **10**, 123-131.

Murray, A. and T. Grubesic (2002). "Identifying non-hierarchical spatial clusters." International Journal of Industrial Engineering 9, 86-95.

Murray, A. and Estivill-Castro, V. (1998). "Cluster discovery techniques for exploratory spatial data analysis." *International Journal of Geographical Information Science* **12**, 431-443.

Murray, A. and Gerrard, R. (1997). "Capacitated service and regional constraints in location-allocation modeling." *Location Science* **5**, 103-118.