Social Indicators Research (2007) 82: 287–309
DOI 10.1007/s11205-006-9034-x

LUC ANSELIN, SANJEEV SRIDHARAN and SUSAN GHOLSTON*

USING EXPLORATORY SPATIAL DATA ANALYSIS TO LEVERAGE SOCIAL INDICATOR DATABASES: THE DISCOVERY OF INTERESTING PATTERNS

(Accepted 29 May 2006)

ABSTRACT. With the proliferation of social indicator databases, the need for powerful techniques to study patterns of change has grown. In this paper, the utility of spatial data analytical methods such as exploratory spatial data analysis (ESDA) is suggested as a means to leverage the information contained in social indicator databases. The principles underlying ESDA are illustrated using a study of clusters and outliers based on data for a child risk scale computed for countries in the state of Virginia. Evidence of spatial clusters of high child risks is obtained along the Southern region of Virginia. The utility of spatial methods for state agencies in monitoring social indicators at various localities is discussed. A six-step framework that integrates spatial analysis of key indicators within a monitoring framework is presented; we argue that such a framework could be useful in enhancing communication between State and local planners.

KEY WORDS: spatial analysis, global association, local association, Community Health Indicators, state-level planning

1. INTRODUCTION

In recent years, a number of efforts have encouraged the collection and monitoring of community-level social indicators on the well-being of children. Typically, these initiatives gather data pertaining to risk and protective factors for children across a number of domains, including community, family, school and individual characteristics. Examples of such initiatives include the Annie E. Casey Foundation's national and statewide Kids Count projects (Annie E. Casey Foundation, 2003; Voices for Virginia's Children, 2005), and Hawkins and Catalano's Communities that Care (Hawkins and

* Given cost considerations, the original color figures have been printed in black and white in this article. The color figures can be obtained from http://www.chs.med.ed.ac.uk/rubhc/evaluation/

Catalano, 1992). Similarly, the Society for Prevention Research has recently set up a study group to focus "on the monitoring of the well-being of children and adolescents as one of its strategic goals" (Society for Prevention Research, 2003). Child indicators provide a natural way of (a) describing the condition of children; (b) setting goals to reduce risk factors and increase protective factors; and (c) monitoring progress toward reaching the goals of improving child outcomes (Land et al., 2001).

Researchers have begun to examine the relationships between these various social indicators (Bramstedt and O'Hare, 2002) and whether some of them might form underlying scales. Examples of such scales include the High-Risk Kids (Kids Count, 1999) and the Index of Child and Youth Well Being (Land et al., 2001). In this paper, we describe how an explicit *spatial* analysis of social indicators can contribute to the discovery of "interesting patterns" in the distribution of child risk and can consequently aid in informing potential policy and program development. By "interesting patterns" we mean areas that have "atypical" values of risks or protective factors (in relation to neighboring areas) or distinct patterns of spatial association (Anselin, 1994, 1999b). We argue that such areas with "interesting patterns" provide opportunities for a centralized planner to "learn" from local areas. We view our focus on "interesting patterns" as a means of reducing the "search-space" for a centralized planner monitoring changes in multiple indicators across a number of communities.[1] In the discussion section, we present a six-step process that summarizes how such a focus on "interesting patterns" can help the centralized planner learn from local areas.

We approach this by implementing an *illustrative* example: we first develop a risk scale for children using principal component analysis techniques;[2] we then examine the spatial distribution of the children's risk scale in 1995 and 2001 with a special focus on identifying significant local clusters and outliers; finally, the policy implications of explicitly accounting for the spatial patterns are discussed.

Our primary motivation is that a large number of data is already being collected at a spatially disaggregated level, such as a county, but most policies are still formulated at a spatially aggregate scale, such as a state. Therefore, emphasizing local and regional disparities by means of methods of spatial analysis enhances and leverages the informational potential of existing databases.

In the empirical illustration, we focus on the Virginia Kids Count database and include only variables that provide complete coverage of the State, are easy to obtain and were of substantive interest to staff involved in community monitoring at the Virginia Department of Juvenile Justice.

Given ongoing efforts at collecting data, similar analyses could be undertaken in any State. We chose the period between 1995 and 2001 to illustrate the importance of space–time analysis and to provide some initial insight into the effect of major policy changes. Nationally, in 1997, Aid to Families with Dependent Children (AFDC) was replaced by a new federal block grant called Temporary Assistance to Needy Families (TANF). Also, Virginia constitutes an interesting case study since in July 1995, the state began phasing in its own Welfare Reform package. This changed eligibility requirements, including a two-year limit on receiving benefits, capping benefits for children born after a family began receiving benefits and urging recipients into the workforce. AFDC/TANF caseloads were reduced by 34% in the first 29 months of Virginia's Welfare Reform. In addition, Virginia underwent Juvenile Justice Reform in 1996. In this illustrative example, we explore the possibility that policy shifts of this scope are associated with changes in spatial patterns of child indicators over time.

We suggest that there is an important role for spatial analysis in monitoring changes in social indicators using local, disaggregated community-level data. Further, we argue that such a focus on spatial patterns of social indicators can assist in developing policies that can be responsive to spatial variations in social problems. All too frequently, program planning has relied upon *aggregated* levels of social indicators across the various jurisdictions yielding policies that apply uniformly across the state. There are problems with such an aggregated approach to state-level policy making and planning: (i) risk indicators might vary dramatically across different regions of a State and (ii) the factors associated with risk might vary across different regions. In the spatial analysis literature, this is referred to as *spatial heterogeneity*. Further, the flow of data usually is from localities to the state. Rarely is there a reverse flow of information *from* the State *to* the localities. Different prevention programs might work in different jurisdictions. Programs designed to meet a statewide need may have little relevance to the needs of specific localities. A clear partnership between state and local levels is more likely to be effective in addressing social issues.

We argue in favor of a system that would: (i) track indicators and the "drivers of indicators"; and (ii) apply a methodology that can systematically explore the dynamic context of social indicators (Sampson, 1993). It is important for such an approach to be sensitive to the fact that explanations of social indicators can vary across different regions. Ideally, such an approach should help communication between the State and local planners, and promote a system of dynamic feedback. Problem behaviors occur *in a* community within the context of *that* community, and the solutions must be sensitive to local needs, assets and people. In the context of our example, Virginia is diverse in geography, economy and the nature of its social problems. The imposition of programs from "central office," is not likely to address the most critical issues equally in every locality and therefore will be hampered in its success.

We follow Anselin (1999a) in distinguishing three important ways that spatial analysis can help enhance the toolbox of the social scientist (see also Goodchild et al., 2000):

(i) *Data integration:* Spatial analysis provides a basis for integration and data collection at different spatial scales and time dimensions; this is especially important in the case of social indicators research because data tends to be collected at a variety of spatial levels and temporal frequencies. Data integration is a central function of the application of Geographic Information Systems (GIS).

(ii) *Exploratory spatial data analysis* (*ESDA*)*:* ESDA is a subset of exploratory data analysis (EDA) that focuses on the distinguishing characteristics of spatial data-specifically on spatial autocorrelation and spatial heterogeneity. More specifically, ESDA is a collection of techniques to describe and visualize spatial distributions, identify atypical locations or spatial outliers, discover patterns of spatial association, clusters or hot spots, and suggest spatial regimes or other forms of spatial heterogeneity (Anselin, 1994, 1999b).

(iii) *Confirmatory spatial data analysis:* Spatial modeling techniques, such as regression analysis can also be implemented to explicitly incorporate the mechanisms underlying the spatial patterns. This encompasses a broad range of activities including model estimation, diagnostics, specification tests, and spatial predictions (Anselin, 1988; Cressie, 1993; Anselin and Bera, 1998).

The focus of the present paper is on the utility of *exploratory spatial data analysis* in uncovering interesting patterns of child risk. We also describe how finding spatial patterns can inform policy development. The move towards an ESDA framework is consistent with the expanding role of GIS in social planning (Page, 1993; Hugo, 1994; Plane and Rogerson, 1994; Kirby and Foldy, 1998).

2. RESEARCH QUESTIONS

We focus on three key questions relating to the spatial patterns of child risk in the illustrative example:

1. *Are risks randomly distributed across Virginia?* If risks are randomly distributed, social policy focused on children, perhaps does not need to explicitly incorporate spatial information into planning efforts. On the other hand, if the underlying spatial distribution of risk is not random, an argument can be made that policy and program development should explicitly incorporate spatial information and be targeted to account for the regional disparities.
2. *Are there spatial clusters or "hot spots" of high-risk communities?* In other words, are counties that have high levels of risk surrounded by counties that also have high levels of risk? Such a focus can help in identifying counties or regions in which there might be greater need for intervention and targeting of resources.
3. *Is there a relationship between spatial patterns of child risk in 1995 and 2001?* Given the importance of the period between these years, it would be especially useful to study the relationships between risks in those years.

3. DATA AND MEASURES

KIDS COUNT in Virginia is a project of the Annie E. Casey Foundation (Casey Foundation, 2003). Its goal is to distribute objective, comparable data on the status of Virginia's children and families to improve the quality of discussion, policy and programs for federal, state and local decision-makers, program providers and advocates. Kids Count is an excellent resource for data as there are Kids Count projects in all states and the social indicators are tailored to the needs of each state.

KIDS COUNT in Virginia annually gathers a variety of social indicator data including measures of delinquency, child abuse and neglect, education, and the economy (Voices for Virginia's Children, 2005). As a wide variety of indicators were available, we initially considered scales of child risk with a number of indicators. However, as the measurement properties[3] of the resultant scales with multiple indicators was predictably complex, given our illustrative goals, we chose the following three measures to construct the child risk scale: *Prenatal Care Rates, Low Birth Weight Rates, and Infant Mortality Rates.* Our choice of measures was driven in part by discussions with research staff involved in community monitoring at the Virginia Department of Juvenile Justice. Each of these measures, lack of early prenatal care, low birth weights, and high infant mortality are indicative of very low levels of community health.

Low Birth Weight Rate is the number of babies weighing under 2500 g (5 pounds, 8 ounces) divided by the total number of live births. Low birth weight babies are more likely to die, suffer developmental disabilities, or have impaired sight or hearing (Voices for Virginia's Children, 2005, p. 28). The data are reported by place of mother's residence, not place of birth. Rates reflect the number of low birth-weight babies per 100 live births. *Infant Mortality Rate is* the number of all children who were under the age of one year at the time of death divided by the total number of live births. The infant mortality rate reflects the overall quality of life in a community. Infant mortality correlates positively with multiple social problems including poverty, unemployment, illiteracy, poor maternal health, poor parenting and lack of access to intensive medical care (Voices for Virginia's Children, 2005, p. 36). Rates represent the number of deaths per 1000 live births. *Prenatal Care Rate is* the number of mothers who saw a physician during the first trimester of pregnancy divided by the total number of live births. Lack of prenatal care increases risks to both mother and fetus and correlates with low birth weight, infant mortality and developmental delays. Poverty and lack of health insurance are the major barriers to prenatal care (Voices for Virginia's Children, 2005, p. 30). Rates represent the number of women seeking early care per 100 live births.[4]

One of the challenges is to create a single scale that provides a summary information on the multiple indicators; As an example, the Kids Count Data book (Annie E. Casey Foundation, 2005) assesses overall rank of child well-being on 10 key indicators of child well-being. The Kids Count index summarizes the information by creating a single overall rank using a standardization process: by creating a standardized score for each indicator; adding the standardized score and then ranking the states based on the summed standardized scores. All indicators get the same weights and no attempt is made to incorporate the relative importance of each indicator in the summed score. Another recent scale developed by Ken Land and colleagues (Land et al., 2001) uses a 28-item index to measure child well-being. The 28 indicators are classified into seven domains of child's well being: material well being, social relationships with family and peers, health, safety and behavior concerns, educational attainments, place in community, emotional and spiritual well-being. The individual time series are indexed by percentage change from a base year 1975 – subsequent annual observations are computed as percentages of the base year values. A summary index is calculated by averaging equally across the individual component series. Given the illustrative focus in this paper, we implement principal component analysis to summarize the information across the three scales. The principal components scale was developed separately for each of the time periods.[5]

The first step in this analysis was to determine whether there is an underlying construct among these factors. If so, the construct can be used to simplify the analysis and allow us to begin to look for spatial patterns in the data.

One of the problems with the calculation of the rates for infant mortality, low birth weight and prenatal care, is that a number of counties have very small populations. In such counties, the calculated "crude" rate may be a poor measure of the underlying risk and spuriously be identified as "outliers." More specifically, the precision of the calculated rate as an estimate of the risk decreases as the denominator decreases (this is referred to as variance instability). A number of transformation and smoothers have been suggested in the literature, particularly in the context of disease mapping and epidemiology to address this problem of small populations (e.g., Clayton and Kaldor, 1987; Marshall, 1991). The solution implemented in this paper is the Empirical Bayes Smoother (EBS). The EBS uses Bayesian principles to guide the adjustment of the crude rate estimate and uses information from the rest of the sample to make the rate adjustments. This principle is referred to as "shrinkage" in the sense that the crude rate is moved (shrunk) towards an overall mean, as an inverse function of the inherent variance. In other words, if a crude rate estimate has a small variance (i.e., is based on a large population at risk), then it will remain essentially unchanged. In contrast, if a crude rate has a large variance (i.e., is based on a small population at risk, as in small area estimation), then it will be "shrunk" towards the overall mean. In a Bayesian approach, the overall mean is a prior, which is conceptualized as a random variable with its own 'prior' distribution (for a recent illustration in a GIS environment, see Anselin et al., 2004).

The approach adopted in the paper is to calculate the principal component scale based either on the crude rates or on the smoothed rates. The results for both sets of components are compared in terms of what they imply for spatial patterns in the data.

4. METHODOLOGY

We briefly describe the principal component analysis, and review the salient ESDA methods used. All spatial analysis was carried out using the GeoDa software package (Anselin et al., 2006).[6]

4.1. *Principal Component Analysis*

Standard principal component analysis models are used to study the relationships between the various indicators and to develop a risk scale. Principal component analysis techniques were implemented to transform the three risk measures into a single continuous scale for county-level child risk. In each year, the transformed risk score had a mean of 0 and a standard deviation of 1. A high positive value on this scale implied a higher level of risk. The principal component scale was developed with both the crude rate and the smoothed rates.

4.2. *Outlier Maps*

An important aspect of ESDA is the visualization of extreme values. In traditional EDA, a standard tool is the box plot (Tukey, 1977), which shows the median and four quartiles, as well as an indication of "extreme" values, defined with respect to the interquartile range.[7] A box map is an extension of the box plot to the map domain (Anselin, 1999b). It is a choropleth map with six categories: four quartiles, as well as special categories (when warranted) for the lower and upper outliers. A box map is an example of an outlier map, which allows for easy identification of both the location and the value of extreme observations.

Often, a choropleth map based on the areal outline of observations, such as county boundaries, is not a good guide to visualize the spatial distribution of a variable when the areal units are highly heterogeneous. For example, in Virginia, several of the "city counties" are barely visible on a standard choropleth map. High or low values for such counties would therefore be hard to distinguish. A cartogram is a technique to focus attention to the magnitude of the variable of interest, rather than the area of a spatial unit. It is a map where the original layout of the areal units is replaced by a layout in which the size of the area is proportional to a given variable. The placement of the circles is such that the original pattern is mimicked as much as possible, both in terms of absolute location as in term of relative location (spatial arrangement). This requires a non-linear optimization routine, such as the cellular automata algorithm outlined in Dorling (1996). An important concept in ESDA is the principle of linking. It means that any observation highlighted in one of the "views" on the data (e.g., the cartogram) is also highlighted in all others. Linking is a fundamental technique in high dimensional data visualization and underlies the exploratory approaches in GeoDa (Buja et al., 1996; Anselin et al., 2006).

4.3. *Global Spatial Autocorrelation*

Central to ESDA is the notion of spatial autocorrelation or spatial association: the phenomenon in which locational similarity (observations in spatial proximity) is matched by value similarity (attribute correlation). Global spatial autocorrelation is a measure of overall *clustering*, and is assessed by means of a test of a null hypothesis of random location. Rejection of this null hypothesis suggests a spatial pattern or spatial structure, which provides additional information about the phenomenon under study. The most familiar test for spatial auto correlation is Moran's I (see Cliff and Ord, 1981; Upton and Fingleton, 1985).[8] This statistic is essentially a cross product correlation measure that incorporates "space" by means of a spatial weights matrix W. Significance can be based on analytical derivations, or, more commonly, on a comparison to a reference distribution obtained by randomly permuting the observed values.

The spatial weights merit some special attention. Each row i of matrix W has elements w_{ij} corresponding to the columns j. The structure of the w_{ij} expresses a prior notion of which locations are important in driving the spatial correlation, in the sense that non-zero values represent "neighbors." Many different perspectives exist on which the values of the w_{ij} can be based. In practice, it is near impossible to choose a "best" weights matrix and typically one assesses the sensitivity of the results to the selection of weights. Common specifications are simple contiguity (sharing a common border), either of a "rook" variety (only pure borders) or of a "queen" variety (both borders and common vertices). These terms are derived from an analogy to a chess board, where the rook neighbors would be the four locations to the North, South, East and West, and the queen neighbors would also include the corner elements (for a total of eight neighbors). Other criteria to construct spatial weights are derived from the distance between the (centroids of) locations. For example, a distance band may be selected such that all locations within the given distances are considered to be "neighbors" (i.e., the corresponding element w_{ij} is non-zero). Alternatively, nearest neighbors may be selected (see Messner and Anselin et al., 2006 or further discussion and examples). In our empirical example, we used five different spatial weights.[9]

An interesting aspect of the Moran's I statistic is that it can be visualized as the slope in a scatter plot of the spatially lagged variable (Wx) on the original variable (x), or a so-called Moran scatter plot (Anselin, 1995). This provides an easy way to categorize the nature of spatial autocorrelation into four types, corresponding to spatial clusters and spatial outliers. Specifically, observations in the lower left (low–low) and upper right (high–high) quadrant represent potential spatial clusters (values surrounded by similar neighbors), whereas observations in the upper left (low–high) and lower right (high–low) suggest potential spatial outliers (values surrounded by dissimilar neighbors). The classification in the Moran scatter plot is only suggestive of clusters or outliers and does not indicate significance. The latter requires a formal hypothesis test (see below).

The Moran scatter plot has recently been extended to the analysis of space–time patterns (Anselin et al., 2002). This implements a notion of bivariate spatial autocorrelation, i.e., the correlation between values at a location at a given point in time and its neighbors at a different point in time. As in the cross-sectional Moran scatter plot, the slope corresponds to the spatial correlation statistic and the four quadrants suggest four different types of association. By changing the time period for the axes both outward diffusion (value at t in location i related to neighbors at time $t + 1$) as well as inward spread (value at $t + 1$ in location i related to neighbors at time t) can be visualized. Significance of the bivariate Moran's I statistic is assessed by means of a permutation approach.

4.4. *Local Spatial Autocorrelation*

Tests for the presence of global spatial autocorrelation only indicate overall clustering, not where the clusters or outliers are located, nor what type of spatial correlation is most important (e.g., correlation between high or between low values). Local indicators of spatial autocorrelation, or LISA address this issue. Specifically, the Local Moran statistic (Anselin, 1995) provides a means to assess significance of "local" spatial patterns. In combination with the classification into four types of association, this indicates significant local clusters (high–high or low–low) or local spatial outliers (high–low or low–high). Significance is typically based on a conditional permutation approach (for details, see Anselin, 1995).[10] A map showing locations with significant Local Moran statistics, classified by type of spatial correlation is referred to as a LISA Cluster map. It is an important tool in identifying "interesting" locations and in assessing the extent to which the spatial distribution exhibits "spatial heterogeneity." In Anselin et al. (2002), the notion of a LISA Cluster map is extended to a bivariate context, which suggests a ready application to space–time local autocorrelation. The resulting bivariate LISA Cluster map shows significant locations classified by the association between the value at one point in time and the value for its neighbors at a different point in time, suggesting possible diffusion patterns.

It is important to keep in mind that these exploratory techniques are only suggestive of possible hypotheses and relations. Confirmation of such patterns belongs to the domain of spatial modeling, which is outside the scope of the current article. The main contribution of the ESDA is to highlight potentially interesting features in the data, and to facilitate the discovery process.

5. RESULTS

The analysis was conducted using both the crude rates and the smoothed rates. In general, the results of the *spatial* analysis were very similar between the two. We will focus primarily on the smoothed rates, but report both when interesting differences merit to be highlighted.[11]

5.1. *Principal Component Analysis*

A single, underlying dimension is extracted from the three indicators of risk using classical principal component analysis. The component loadings are listed in Table I. The resulting scale explains close to 60% of the variation in the social indicators in 1995, and 56% of the variation in 2001. The underlying factors load reasonably strongly with all three measures for both 1995 and 2001, with only minor differences between the two years.[12] This provides support to reduce the three variables to a single child risk dimension for further analysis.

5.2. *Outliers*

A first look at the spatial distribution of the child risk factor across Virginia counties is provided by the "box maps" listed in Figures 1–4. Figures 1 and 2 illustrate the spatial pattern for the scores based on crude rates in 1995 and 2001. In 1995, there were four upper outliers (Brunswick, Richmond,

TABLE I

Component Loadings, Child Risk Scale

	1995	2001
Infant mortality rate	0.83	0.80
Prenatal care rate	−0.69	−0.67
Low birth weight rates	0.80	0.77

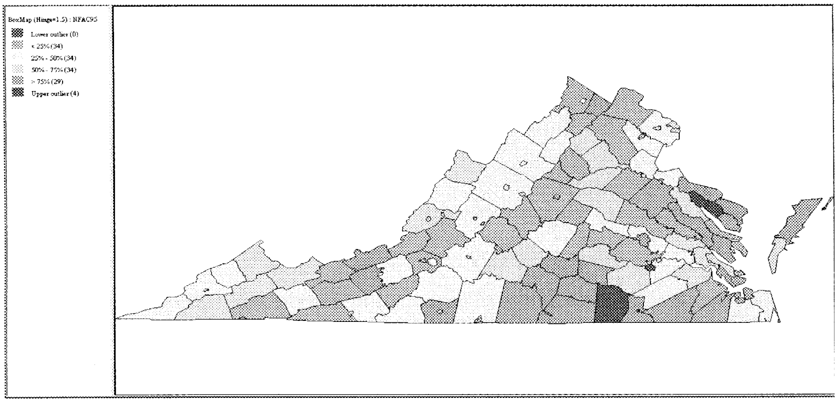

Fig. 1. Box map for risk factor (based on crude rates), 1995.

Petersburg and Emporia), in 2001, six (Richmond, Emporia, and Northumberland, Greensville, Westmoreland and Charles City), only two of which were common between the two years. Note that several of the "outliers" are counties with some of the smallest population sizes in the state, suggesting a potentially misleading classification as a spurious outlier due to extra variance instability. Also note that some of these outliers are difficult to distinguish on the county choropleth map, due to the many "city counties" in Virginia. For example, Emporia, which is an outlier in both years, is a tiny speck inside the much larger Greensville county on the Southern border of the state, barely distinguishable in Figures 1 and 2.

In Figures 3 and 4 the box map is shown for the risk factors computed from smoothed rates, and by "linking" the choropleth map to a cartogram. Figures 3 and 4 highlight the differences between the crude and smoothed rates (compare to Figures 1 and 2). Whereas the overall spatial pattern is fairly similar (higher risk along the Southern border, lower risk around Washington DC), there are some interesting differences in the outliers. First, there is much greater agreement between the two years, with Norfolk, Richmond City, Portsmouth and Danville identified as outliers at both points in time. Second, there is very little agreement with the outliers identified for the crude rates, highlighting the small area problem. Overall, the outlier analysis would suggest that some counties show persistent elevated risk factors over time, distinct from the overall distribution exhibited by the other counties.

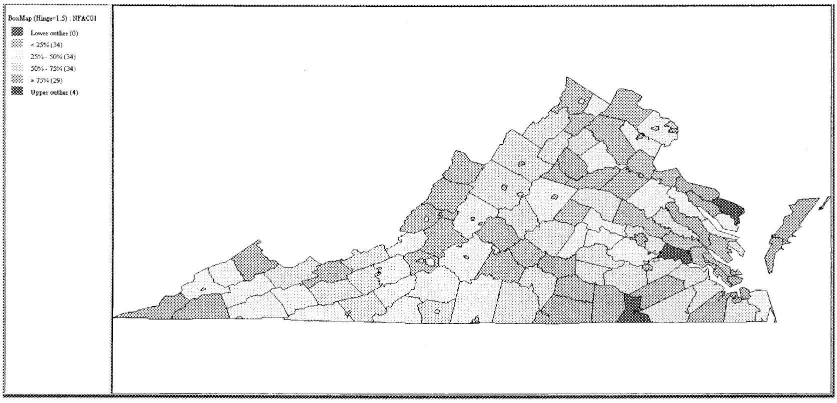

Fig. 2. Box map for risk factor (based on crude rates), 2001.

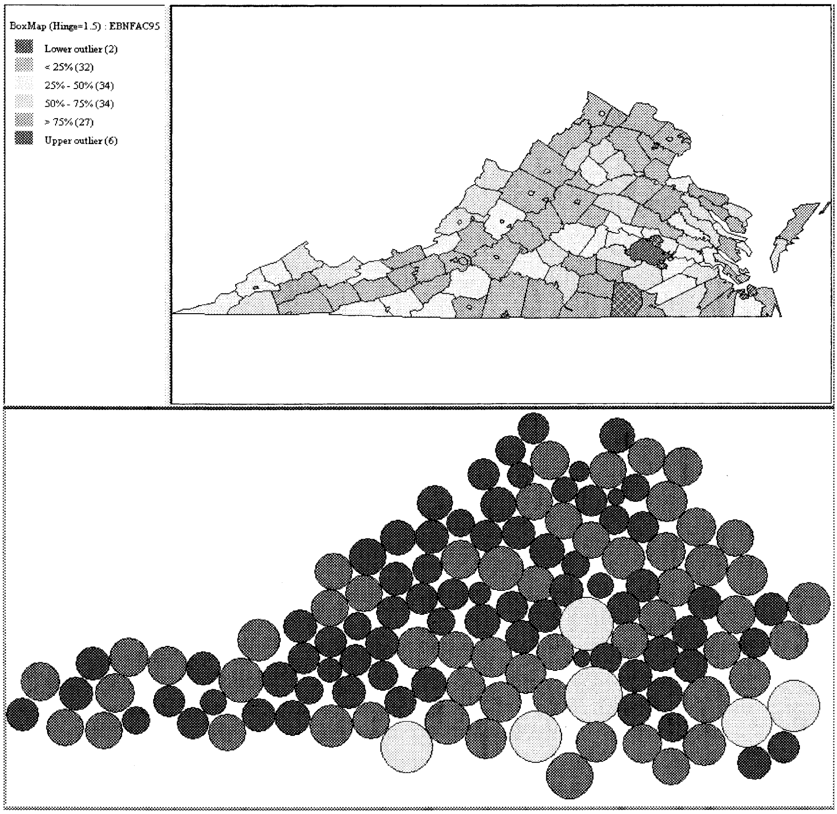

Fig. 3. Box map with linked cartogram with outliers for risk factors (based on EB smoothed rates), 1995.

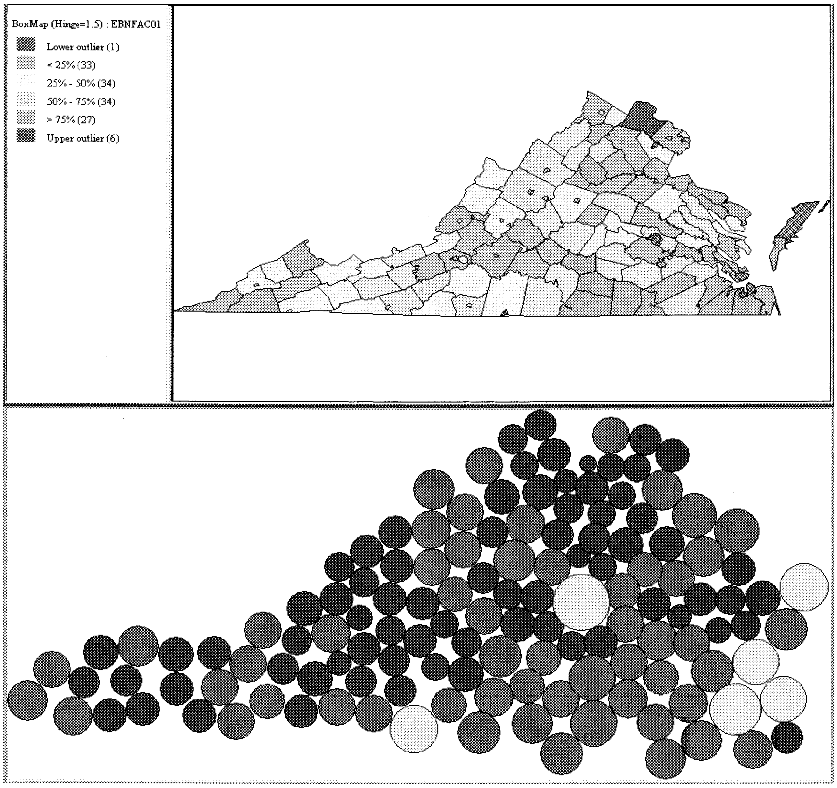

Fig. 4. Box map with linked cartogram with outliers for risk factors (based on EB smoothed rates), 2001.

5.3. *Global Association*

Our second focus of attention is the extent to which the spatial patterns exhibited are compatible with a notion of randomness (the null hypothesis), or instead show a significant spatial structure. We refer to this as testing for *global* association, since it pertains to the overall pattern in the data, rather than localized features. The results for Moran's I test for spatial autocorrelation are given in Table II, with pseudo significance values (based on a permutation approach) listed in parentheses. We consider risk factors computed from both the crude and smoothed rates and use five different spatial weights matrices to assess the extent to which the indication of spatial autocorrelation depends on the assumed spatial structure. The results suggest significant and positive spatial autocorrelation across all spatial weights, except for the smoothed factor in 1995. In that case, the

TABLE II

Moran's I statistic for spatial autocorrelation in risk factors (p-value in parentheses)

Spatial weights	Risk 95	EB Risk 95	Risk 01	EB Risk 01
Rook	0.2001	0.0475	0.3580	0.2471
Contiguity	(.003)	(.200)	(.001)	(.001)
Queen	0.2102	0.0837	0.3494	0.2586
Contiguity	(.003)	(.092)	(.001)	(.001)
Nearest	0.1658	0.0321	0.2737	0.2237
Neighbors (4)	(.006)	(.240)	(.001)	(.003)
Nearest	0.1650	0.0828	0.2641	0.2182
Neighbors (6)	(.002)	(.043)	(.001)	(.001)
Distance band	0.2026	0.1362	0.2835	0.2420
(35mi)	(.001)	(.004)	(.001)	(.001)

values of the spatial autocorrelation statistic are much lower and only significant for two of the five weights (and only weakly so, for the six nearest neighbor weights). In 2001, there are only minor differences between the crude and smoothed factors, consistently indicating a significant degree of overall clustering, rather than a random pattern. The systematic variation of risk factors across space would suggest that this spatial pattern should be taken into account.

A consideration of the space–time association is visualized in the space–time Moran scatter plot matrix shown in Figure 5. This figure contains three graphs, two of which pertain to space–time association and one to the usual serial (time) association.[13] The center graph suggests a strong degree of positive correlation between the risk factors in 2001 and in 1995 (correlation coefficient of 0.73). The scatter plot is drawn for standardized values, with each of the quadrants showing which counties stayed below (above) average

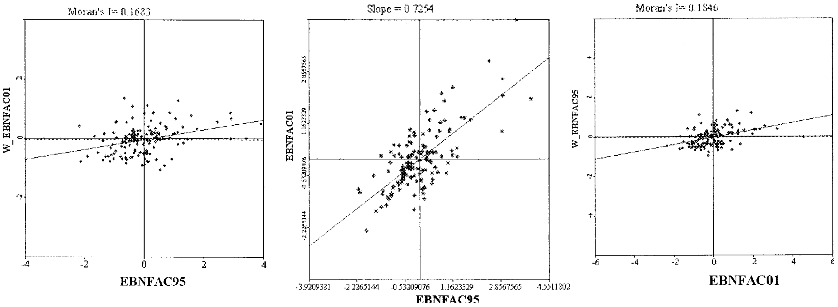

Fig. 5. Correlation and space–time correlation between risk factors (smoothed) in 1995 and 2001.

(the lower left and upper right quadrants) and which counties "moved" from below to above average risk (upper left) and vice versa (lower right). In the graph on the left of the panel, the risk factor in 1995 is related to the average of the "neighbors" (as defined by a specific spatial weights matrix) in 2001. The graph on the right shows the reverse, the correlation between risk in a county in 2001 and its neighbors average risk in 1995. Both of these provide a way to quantify possible "diffusion" (from the center out in the left hand figure, and from the outer ring in on the right hand side). The corresponding measures of space–time autocorrelation (0.17 and 0.18, respectively), while weaker than pure cross-sectional spatial autocorrelation, are highly significant ($p < 0.001$), confirming the presence of significant spatial patterning over time.

5.4. *Local Association*

The previous analysis suggested non-randomness in the overall spatial pattern. A focus on *where* this non-randomness may be located, in terms of significant clusters or spatial outliers (as opposed to overall *clustering)* is provided by an analysis of the local indicators of spatial association (LISA). Figures 6 and 7 are so-called LISA Cluster maps for the risk factors in both years. The highlighted locations are "significant" (at $p < 0.05$) local clusters or outliers, based on the Local Moran statistic.[14] Four categories of spatial autocorrelation are distinguished, two of which suggest clusters and two which suggest outliers.[15] High values surrounded by high values, and low values surrounded by low are clusters, whereas high values surrounded by

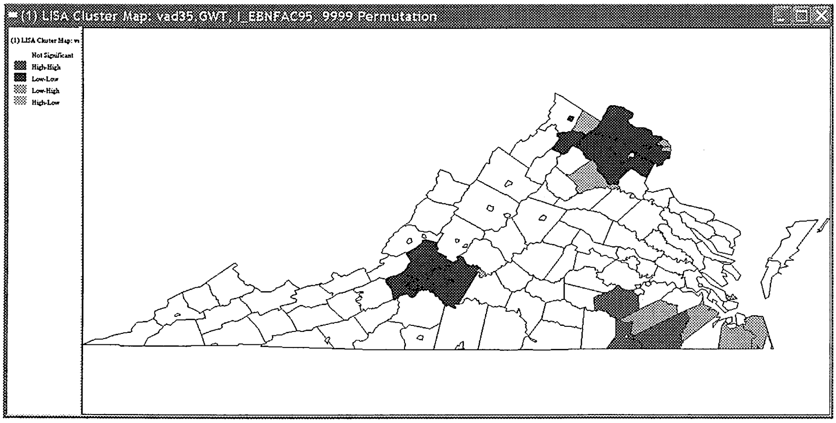

Fig. 6. LISA cluster map for risk factor (smoothed rates), 1995.

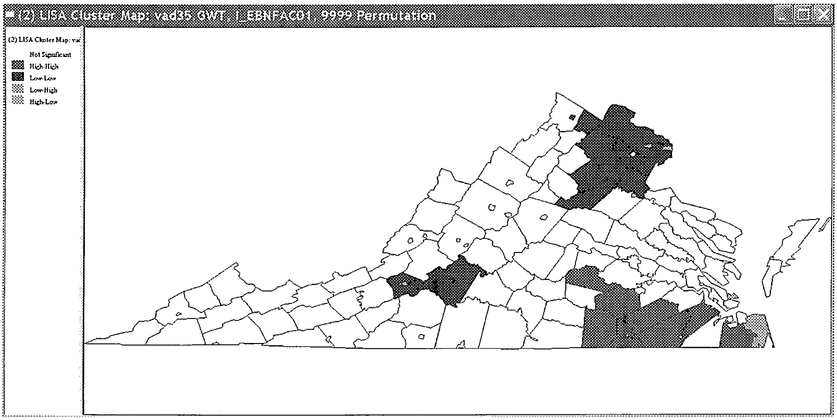

Fig. 7. LISA cluster map for risk factor (smoothed rates), 2001.

low values and low values surrounded by high values are spatial outliers, indicating locations where the pattern changes significantly. The overall pattern in both maps is similar, showing a cluster of low risk counties in the North Eastern corner of the state (around Washington, DC) and a cluster of high-risk counties along Eastern edge of the Southern border. Interestingly, there is also a distinct cluster of low risk around Roanoke and Bedford counties in Western Virginia.

The low risk is associated with an urban area characterized by dense population, diverse ethnicity and socioeconomic status, and some of the highest per capita incomes in the nation. The cost of living is high and is driven by a thriving economy. There are numerous job opportunities from the service industry through highly skilled technical industries. In contrast, the Southern counties are part of a rural area of Virginia, characterized by sparse population, a predominately black lower class, a white middle class, and a relatively small upper class. While there are some factories, the primary industry historically has been farming of tobacco, peanuts and cotton.

The clusters are relatively stable over time, with some indication that both are growing in spatial extent, suggesting a higher degree of spatial polarization over time. For example, several of the high–low outliers in 1995 become swallowed up in the Northern low–low cluster. Similarly, several of the low–high spatial outliers in 1995 become part of the high–high cluster in 2001. This provides evidence of systematic spatial heterogeneity, suggesting that different processes may be at work in different subregions of the state.

A final look is provided by Figure 8, where the clusters and outliers are shown using a space–time Local Moran statistic (relating the value at a

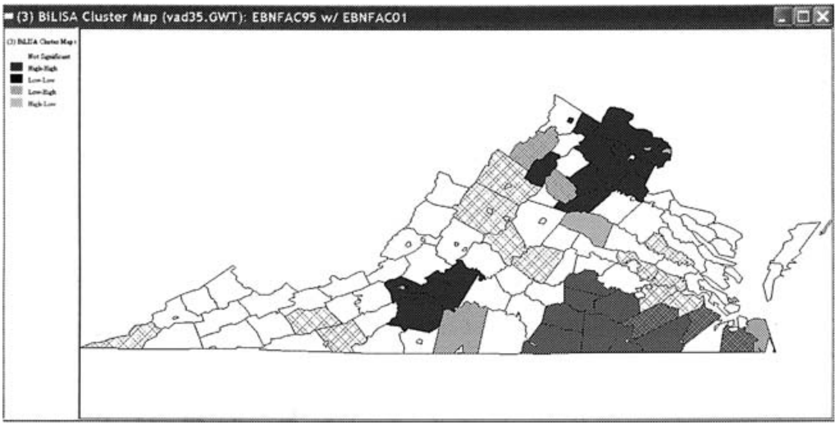

Fig. 8. (Space–time LISA cluster map (smoothed rates) with transitional counties (low to high) highlighted.)

location at one point in time to that of its neighbors at a different point in time). Superimposed on the significant locations are those counties that moved from below average to above average over the time period. It is interesting to note how four of these also become part of significant locations: three are part of the high-risk cluster along the Southern border, and two are high–low outliers in the North. Again, this suggests a growing polarization over time, of high-risk counties along the Southern border of the state and low risk counties in the North.

6. DISCUSSION AND POLICY IMPLICATIONS

The exploration of spatial patterns in child risks in Virginia counties clearly demonstrated the presence of significant spatial clusters of high and low values, as well as some interesting spatial outliers. The clusters and outliers are regionally defined and suggest that an explicit spatial focus on the causal agents may be warranted. We also found that the clusters persisted over time and some indication of a growing of the polarization of the state between low risks near the Washington, DC metro area, as well as in the Western center around Roanoke, and high-risks along the Southern border.

The strong spatial imprint of child risk suggests that an explicit incorporation of spatial information into policy development could yield great benefits. It also highlights that a "one-size-fits-all" state-wide approach to respond to children's risk may not be the most effective in improving child well-being. Our basic contention is that spatial analysis can serve as a useful tool by State agencies to monitor social indicators. Key in our view is a focus on “social learning” from local patterns of child risks. Fundamental to improved learning is a better dialogue between State planners and local administrators.

We outline a six-step process in Figure 9 to help to build a dialogue between State and local planners towards a more effective spatially inspired social policy. The six steps consist of:

(i) *Monitoring:* Developing a Geographical Information System to track county-level measures of child risk over time.
(ii) *Understanding context:* Implementing ESDA to study spatial properties of child risk measures.
(iii) *Identification:* Using exploratory spatial models to identify counties with “interesting patterns” of child risks scale – specifically identify counties with “problem” or “exemplary” levels of risk.
(iv) *Communication:* Policymakers and state planners communicate with local planners in the counties with “problem” or “exemplary” levels of risk to better understand the factors associated with the underlying risk.
(v) *Diagnosis:* State planners work closely with local planners to develop locally driven explanations for “problem” or “exemplary” levels of risk.
(vi) *Response/Learning:* State and local planners develop programs for “problem” counties and incorporate lessons learned from “exemplary” counties into State Planning activities.

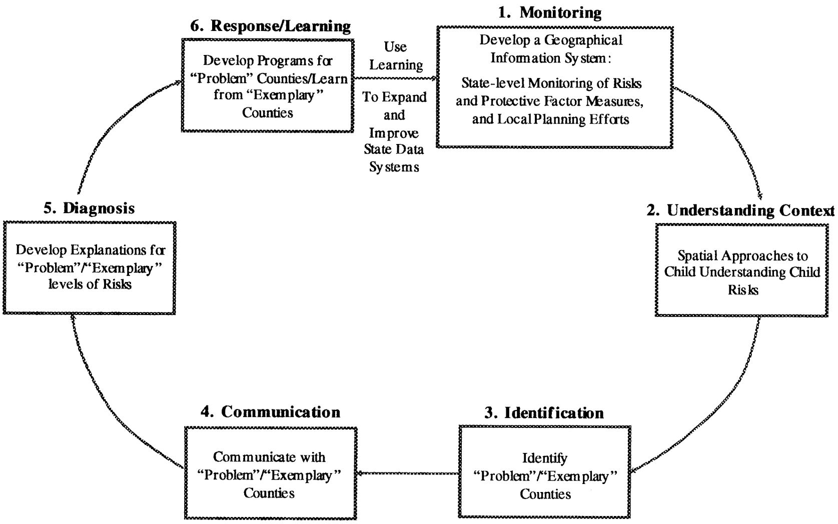
1. Monitoring
Develop a Geographical Information System:
State-level Monitoring of Risks and Protective Factor Measures, and Local Planning Efforts
2. Understanding Context
Spatial Approaches to Child Understanding Child Risks
3. Identification
Identify “Problem”/“Exemplary” Counties
4. Communication
Communicate with “Problem”/“Exemplary” Counties
5. Diagnosis
Develop Explanations for “Problem”/“Exemplary” levels of Risks
6. Response/Learning
Develop Programs for “Problem” Counties/Learn from “Exemplary” Counties
Use Learning
To Expand and Improve State Data Systems

Fig. 9. Towards a dynamic system of comprehensive state planning.

One limitation of this paper is that the child risk scale was not substantively driven but only intended to provide an illustrative example of the utility of spatial analysis in monitoring social indicators. The primary focus of this paper has been on how a spatial approach can help with identification of "interesting patterns." The principal components scale was intended to be a convenient means of summarizing the data with multiple indicators. The principal components analysis was done with both the crude rates and the smoothed rates and the results were found to be comparable. The spatial approaches discussed in this paper can also be applied directly to each social indicator or other index of child risks such as the High-Risk Kids (Kids Count, 1999) and the Index of Child and Youth Well Being (Land et al., 2001). The analysis conducted with the principal component analysis was also conducted with each of the measures *Prenatal Care Rates, Low Birth Weight Rates*, and *Infant Mortality Rates*. Spatial patterns were especially strong in the Prenatal Care Rates and Low Birth Weight Rates in both 1995 and 2001. Weak spatial patterns are observed for Infant Mortality Rates. Spatial polarization between 1995 and 2001 was especially strong for the Low Birth Weight Rate measure. Given our illustrative focus, we do not explore the substantive implications of the differences in spatial imprints in the three measures.

A second potential limitation in our illustrative example is the use of county-level measures to build a monitoring system. From a theoretical point of view, a finer level of analysis such as census tract might capture features of a community more clearly. However, our choice of counties as the unit of analysis is driven by policy considerations. The Virginia Department of Juvenile Justice is organized at the local level by counties; therefore, an understanding of the county-level patterns of social indicators may be useful in helping the Department of Juvenile Justice target programs that are specific to each county.

7. CONCLUSIONS

The spatial analytical methods presented in this paper can be useful in maximizing the *informational potential* of social indicator databases that are routinely collected such as the Virginia Kids Count project (Voices for Virginia's Children, 2003). They provide a basis for a spatially explicit social policy that may be able to address the needs of individual communities in a more effective manner. As the user friendliness of tools and software continues to improve, and the databases such as Virginia Kids Count continue to proliferate, there will be increasing opportunities to incorporate spatial analytical tools into State-level monitoring activities.

NOTES

[1] In our experience in the United States, analysts based in State agencies can often have access to a large number of indicators for a number of communities. As an example, an analyst examining changes in 10 indicators in 135 communities over 5 points of time will have 6750 data points to examine and an even larger number of comparisons to make over time. The spatial methods described in this paper provide *one approach* to reducing the analysts "search-space."

[2] Given our focus on an illustrative example, the scale while informed by the literature is not intended to be substantively driven. This illustrative example primarily serves to highlight the key features of our proposed approach.

[3] Specifically the factor structure of the resultant scales.

[4] Statistical information on these measures can be obtained from the Virginia Center for Health Statistics at http://www.vdh.state.va.us/HealthStats/stats.asp.

[5] We reiterate that our focus in this paper is on the utility of spatial analysis for monitoring social indicators – the principal components analysis is only a convenient device to create a composite measure of the three measures. Alternative methods of creating composite summary measures could also have been adopted.

[6] GeoDa version 0.9.5-i was used. The software is available for free downloading from https://geoda.uiuc.edu.

[7] Outliers are values that are outside a "fence," which is computed by adding or subtracting 1.5 times the interquartile range (the difference between the 75 percentile and the 25 percentile) to the third quartile (from the first quartile). This is similar to the 2 standard deviation rule of thumb often used in a parametric context.

[8] The global measure of Moran's I is defined as:

$$I = \sum_i \sum_j w_{ij} \cdot (x_i - \mu) \cdot (x_j - \mu) / \sum_i (x_i - \mu)^2$$

where w_{ij} is the row-standardized contiguity matrix, x_i is the risk scale measure at county i, and x_j is the risk scale measure at county j, and μ is the average level of risk.

[9] These matrices included the rook criteria, queen criteria, distance-based contiguity with distance between center less than 35 miles, and four and six nearest neighbors. All weights were constructed using the GeoDa Tools.

[10] The local measure of Moran's I is defined as:

$$I = \frac{(x_i - \mu)}{\sum (x_i - \mu)^2} \sum_j w_{ij} \cdot (x_j - \mu)$$

[11] The correlations between the principal component scales using the crude and smoothed measures were 0.89 in 1995 ($p < 0.001$) and 0.85 in 2001 ($p < 0.001$).

[12] One of Virginia's cities, Clifton Forge was incorporated into neighboring Alleghany county in 1990s. Data were missing on a number of indicators from Clifton Forge. For both the principal components scale and the ESDA applications, information from Alleghany County was used for Clifton Forge.

[13] We report only the results using the distance band spatial weights and the risk factors computed from smoothed rates. Results for other weights and risk factors computed from crude rates are qualitatively similar.

[14] The Local Moran statistic used in these LISA maps is based on a distance band spatial weights matrix. Results for the other spatial weights are qualitatively similar.

[15] The distinctions between the low–low and the high–high categories are especially unclear in the printed black and white maps in the map. All of the figures in this paper are available in color at http://www.chs.med.ed.ac.uk/ruhbc/evaluation/

ACKNOWLEDGMENTS

Sridharan's research was supported in part by a National Institute of Justice Grant 2002-IJ-CX-0010 to Westat. Anselin's research was supported in part, by US National Science Foundation Grant BCS-9978058 to the Center for Spatially Integrated Social Science (CSISS).

REFERENCES

Annie E. Casey Foundation: 2003, Kids Count (Annie E Casey, Baltimore, MD).

Annie E. Casey Foundation: 2005, Kids Count Data Book (Annie E Casey, Baltimore, MD).

Anselin, L.: 1988, Spatial Econometrics: Methods and Models (Kluwer Academic, Dodrecht).

Anselin, L.: 1994, 'Exploratory spatial data analysis and geographic information systems', in M. Painho (ed.), New Tools for Spatial Analysis (Eurostat, Luxembourg), pp. 45–54.

Anselin, L.: 1995, 'Local indicators of spatial association-LISA', Geographical Analysis 27, pp. 93–115.

Anselin, L.: 1999a, 'The future of spatial analysis in the social sciences', Geographical Information Sciences 5, pp. 67–76.

Anselin, L.: 1999b, 'Interactive techniques and exploratory spatial data analysis', in P. Longley, M. Goodchild, D. Maguire and D. Rhind (eds.), Geographical Information Systems: Principles, Techniques, Management and Applications (Wiley, New York), pp. 251–264.

Anselin, L. and A. Bera: 1998, 'Spatial dependence in linear regression models with an introduction to spatial econometrics', in A. Ullah and D. Giles (eds.), Handbook of Applied Economic Statistics (Marcel Dekker, New York).

Anselin, L., Y.-W. Kim and I. Syabri: 2004, 'Web-based analytical tools for the exploration of spatial data', Journal of Geographical Systems 6, pp. 197–218.

Anselin, L., I. Syabri and Y. Kho: 2006, 'GeoDa: An Introduction to Spatial Data Analysis', Geographical Analysis 38, pp. 5–22.

Anselin, L., I. Syabri and O. Smirnov: 2002, 'Visualizing multivariate spatial correlation with dynamically linked windows', in L. Anselin and S. Rey (eds.), New Tools for Spatial Data Analysis: Proceedings of a Workshop (Center for Spatially Integrated Social Science, Santa Barbara).

Bramstedt, N. and W. O'Hare: 2002, 'Examining Inter-Relationships Among State-level Measures of Child Well-Being', (Annie E Casey, Baltimore, MD). www.aecf.org/kidscount/sda final 1.1 18.pdf.

Buja, A., D. Cook and D Swayne: 1996, 'Interactive high dimensional data visualization', Journal of Computational and Graphical Statistics 5, pp. 78–99.

Clayton, D.G. and J Kaldor: 1987, 'Empirical Bayes estimates of age-standardized relative risks for use in disease mapping', Biometrics 43, pp. 671–691.

Cliff, A.D. and J.K. Ord: 1981, Spatial Processes, Models and Applications (Pion, London).

Cressie, N.: 1993, Statistics for Spatial Data (Wiley, New York).

Dorling, D.: 1996, Area Cartograms: Their Use and Creation. CATMOG 59 (Institute of British Geographers).

Goodchild, M., L. Anselin, R. Appelbaum and B Harthorn: 2000, 'Towards spatially integrated social science', International Regional Science Review 23, pp. 139–159.

Hawkins, J. and R. Catalano: 1992, Communities That Care: Action for Drug Abuse Prevention (Jossey-Bass, San Francisco).

Hugo, G: 1994, 'GIS & socio-economics', GIS User 6, pp. 46–47.

Kids Count: 1999, 1999 KIDS COUNT Online. (Annie E Casey, Baltimore, MD). http://www.aecf.org/kidscount/kc 1999/overview.htm.

Kirby, R. and S. Foldy: 1998, 'The role of geographic information systems in population health', in R.Williams, M. Howie, C. Lee and W. Henriques (eds.), Geographic Information Systems in Public Health: Proceedings of the Third National Conference, (Centers for Disease Control, Atlanta, pp. 579–587.

Land, K., V. Lamb and S Mustillo: 2001, 'Child and youth well-being in the United States, 1975–1998: Some findings from a new index', Social Indicators Research 56, pp. 241–318.

Marshall, R.: 1991, 'Mapping disease and mortality rates using empirical Bayes estimators', Applied Statistics 40, pp. 283–294.

Messner, S. and L. Anselin: 2004, 'Spatial analyses of homicide with areal data', in M. Goodchild and D. Janelle (eds.), Spatially Integrated Social Science (Oxford University Press, New York), pp. 127–144.

Page, P.: 1993. 'GIS & social sciences', in Proceedings of the Thirteenth Annual ESRI Conference, Vol. I (ESRI, Los Angeles), pp. 385–396.

Plane, D. and P. Rogerson: 1994, The Geographical Analysis of Population with Applications to Planning & Business (John Wiley & Sons, New York).

Sampson, R.: 1993, 'Linking time & place: Dynamic contextualism and the future of criminological inquiry', Journal of Research in Crime & Delinquency 30, pp. 426–444.

Society for Prevention Research (2003) Community Level Monitoring. http://www.preventionresearch.org/commlmon.php.

Tukey, J.: 1977, Exploratory Data Analysis (Addison-Wesley, Reading).

Upton, G. and B. Fingleton: 1985, Spatial Data Analysis by Example (Wiley, New York).

Voices for Virginia's Children: 2005, Virginia Kids Count Data Book (Voices for Virginia's Children, Richmond, VA).

The Evaluation Programme, Research Unit in Health
Behaviour and Change Community Health Sciences, RUHBC
The University of Edinburgh, Medical School
Teviot Place, Edinburgh, EH8 9AG, UK
Sanjeev.Sridharan@ed.ac.uk
S. Sridharan

University of Illinois at Urbana–Champaign and NCOVR
61801, IL, USA
L. Anselin

Virginia Department of Juvenile Justice
Richmond, VA, USA
S. Gholston