

Spatial Statistics Using the SAS Bridge for ESRI

Patricia B. Cerrito, Carol Hanchette, William Dakan, George Barnes, Robert Forbes, University of Louisville, Louisville, KY

ABSTRACT

The SAS Institute, Inc. has developed a bridge to ArcGIS software, developed by ESRI, Inc. ArcGIS is used worldwide for spatial data analysis and, increasingly, for spatial statistics. The bridge allows users to exchange spatial attribute data between SAS and ArcGIS and to import SAS programming directly into ArcGIS. The bridge provides ArcGIS users access to a wide range of statistical functions and enhances the use of spatial statistics to analyze map data. This presentation will demonstrate the effective use of the SAS Bridge for ESRI, using kernel density estimation as an example.

In recent years, health researchers have shown much interest in using kernel density estimation for the exploration and display of point health data. Kernel density estimation provides an alternative to choropleth mapping (e.g. shaded mapping of administrative units) and some of its drawbacks, such as the modifiable area unit problem. Using point health data, this presentation will demonstrate the use of kernel density functions in both ArcGIS and SAS to generate information about the density of health events. The Spatial Analyst module of ArcGIS will be used to represent and map the outcomes as a continuous field.

INTRODUCTION

The University of Louisville received funding from the National Science Foundation to develop a 2-course graduate sequence that was interdisciplinary, combining geographic information systems and SAS applied to public health data. Faculty in the Departments of Geography/Geosciences and Mathematics cross-trained in the use of ArcGIS and SAS (via Enterprise Guide). Faculty also attend webcasts and seminars concerning the SAS Bridge, and the latest upgrades to both ArcGIS and SAS (version 9 in both cases). On the basis of this training, the involved faculty then constructed a syllabus (included in the appendix) for the graduate course, and started to develop research projects for the students enrolled in the course. In addition, faculty began to compare statistical techniques in both geography and statistics to see how the data could best be analyzed.

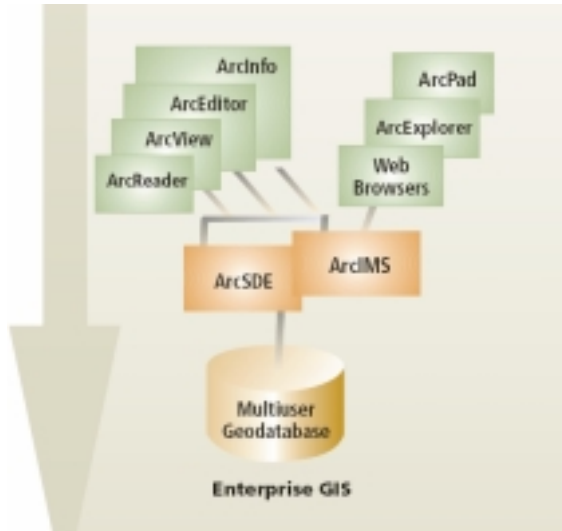
We discovered through the cross-disciplinary discussions that similar terminology conveyed different concepts, and that similar concepts were represented in somewhat different ways. For example, to the statistician, clustering means a group of techniques to define inherent groups in the data observations. To the geographer, clustering means a concentration of specific objects, or illnesses in a particular geographic region. Hypothesis testing is performed in the presence of spatial auto-correlation. To the statistician, this means a mixed models procedure. The statistician can also provide a level of abstraction to the use of statistical methodology. PROC MIXED can be used to examine these data. The geographers define a vector of weights to represent spatial auto-correlation. While these weights are not specifically in PROC MIXED, the mixed models algorithm used does allow a default unstructured, estimated covariance matrix.

One important technique frequently used in geography is kernel density estimation. The statistician uses it as a general tool for any continuous variable but the geographer defines it exclusively as a measurement of distance, with the bandwidth defining a specific distance interval. PROC KDE in SAS can be used to define kernel density estimators.

THE SAS BRIDGE TO ESRI

SAS and ESRI have collaborated to enable data to pass between the two programs. In particular, it is possible to add a SAS stored program directly into ArcGIS. Information about the mapping products can be found at www.esri.com. Information about the SAS Bridge to Esri is available at http://support.sas.com/rnd/datavisualization/BridgeForESRI/V1/stored_process.htm. In order to use the bridge, ArcInfo and SAS must both be installed on the same machine. The mapping tools available are given in Figure 1.

Figure 1. Mapping Products Available From ESRI



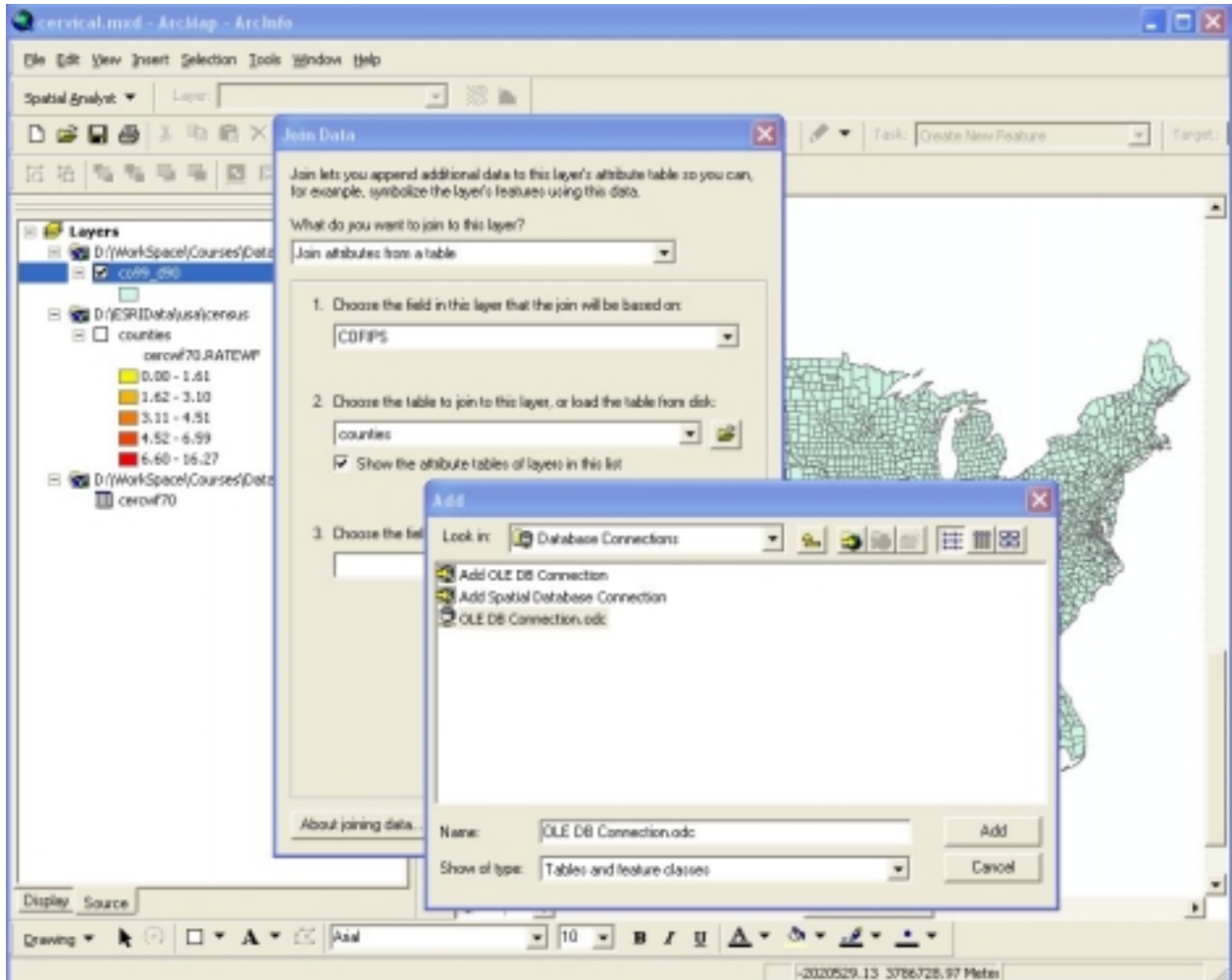
There is a hierarchy in the products. ArcReader allows the user to view maps only. ArcView contains several programs, including ArcMap that allow the construction of maps. In order to use the SAS Bridge, ArcInfo must be available.

After it is installed, an add-in is placed in ArcGIS so that a map layer can be automatically saved as a SAS dataset for direct importation into SAS. At the same time, SAS code can be imported directly into ArcMap and executed directly into ArcMap.

Attribute data from SAS can also be joined with an ArcGIS shapefile via the Join function. Following is an example using a U.S. county shapefile (source: Census TIGER/Line 2000) and cervical cancer mortality data from 1970-1994 (source: National Cancer Institute, *Atlas of Cancer Mortality in the United States 1950-1994*, available at: [HTTP://WWW3.CANCER.GOV/ATLASPLUS/](http://www3.cancer.gov/atlasplus/)). To join a SAS data set to an ArcMap layer:

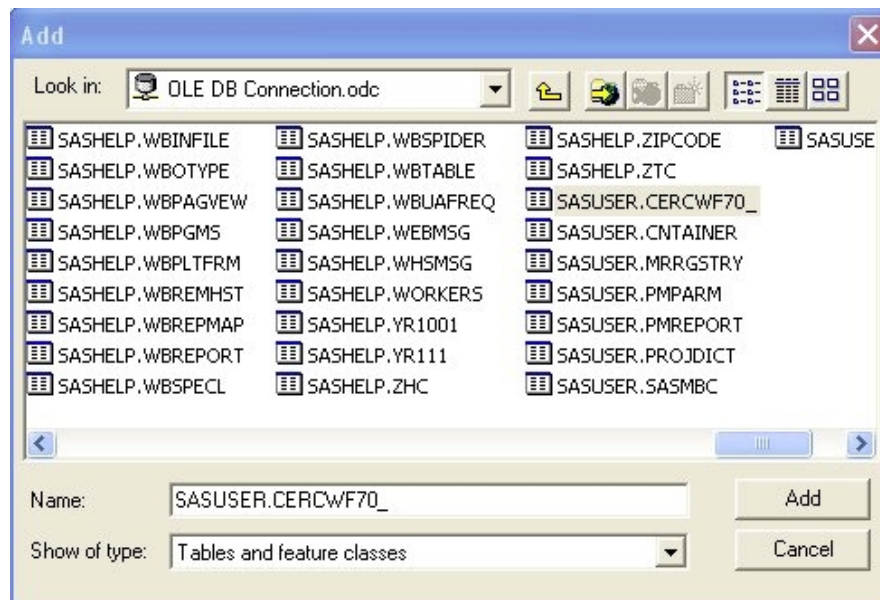
1. Select a layer or shapefile in ArcMap and right-click.
2. Select **Joins and Relates**→**Join** from the pop-up menu.
3. Select **Join attributes from a table**.
4. Choose a field from the ArcGIS layer for the join.
5. To choose a table to join to the layer, hit the browse button (file folder icon) and click the **Database Connections** button on the Join Data window. If an OLE DB connection is not present, one will have to be added.
6. To add an OLE DB Connection:
 - a. Double-click **Add OLE DB Connection** in the Add Data window.
 - b. Select **SAS IOM Data Provider 9.0** from the Provider tab of the Data Link Properties window.
 - c. Click the **Connections** tab and specify **_local_** in the Data Source field.
 - d. Click **OK**.
7. Now select **OLE DB Connection.odc** from the drop-down list of sources. Figure 2 shows an ArcMap session joining data from the U.S. counties map layer. The Join Data and Add Data windows are open. The OLE DB Connection.odc is highlighted in the Add Data window.

Figure 2. ArcMap session



8. Next, select the SAS data set to be added. In Figure 3, the CERCWF70 data set is highlighted.

Figure 3. SAS Datasets Viewed in ArcMap



9. Click **Add**.
10. In the Join Data window, select the corresponding field from the SAS data set for the join.
11. Click **OK**.

For GIS users unfamiliar with SAS libraries, the trickiest part of the join may be determining where the SAS data sets are stored. One bug we have encountered, at this time, is that the formatting of some of the columns is changed when the join occurs. At this time, we are working to resolve this issue.

KERNEL DENSITY ESTIMATION

Kernel density estimation is used in geography as a method to find spatial patterns of point health data. According to Cromley and McLafferty¹, “[the] kernel is moved across the study area, and the density of events is computed within this window. Typically, this window is a circle with a constant radius, or bandwidth. Events within the window are weighted according to their distance from the center of the window, the point at which density is being estimated.” However, it is also defined in the more traditional statistical fashion.²

The kernel density estimate is defined by the equation:³

$$\hat{f}(x) = \frac{1}{na_n} \sum_{j=1}^n K\left(\frac{x - X_j}{a_n}\right)$$

where n is the sample size, K is a known density function, and a_n is a constant depending upon the size of the sample that controls the amount of smoothing in the estimate. Note that for most standard density functions K , where x is far in magnitude from any point X_j , the value of will be very small. Where many data points cluster together, the value of the density function will be high because the sum of $x - X_j$ will be small and the probability defined by the kernel function will be large. However, where there are only scattered points, the value will be small. K can be the standard normal density, or the uniform density. Simulation studies have demonstrated that the value of K has very limited impact on the value of the density estimate. It is the value of the bandwidth, a_n , that has substantial impact on the value of the density estimate. The true value of this bandwidth must be estimated, and there are several methods available to optimize this estimate.

By default, the standard normal density is used for K . However, there are several different methods can be used to estimate the bandwidth as discussed below. The default for the univariate smoothing is that of Sheather-Jones plug in (SJPI):

$$h = C_3 \left\{ \int f''(x)^2 dx, \int f'''(x)^2 dx \right\} C_4 (K) h^{5/7}$$

where C_3 and C_4 are appropriate functionals. The unknown values depending upon the density function $f(x)$ are estimated with bandwidths chosen by reference to a parametric family such as the Gaussian as provided in Silverman:

$$\int f''(x)^2 dx = \sigma^{-5} \int \phi''(x)^2 dx \approx 0.212 \sigma^{-5}$$

However, the procedure uses a different estimator, the simple normal reference (SNR), as the default for the bivariate estimator:

$$h = \hat{\sigma} \left[\frac{4}{(3n)} \right]^{1/5}$$

along with Silverman's rule of thumb (SROT):

$$h = 0.9 \min[\hat{\sigma}, (Q_1 - Q_3) / 1.34] n^{-1/5}$$

and the oversmoothed method (OS):

$$h = 3\hat{\sigma} \left[\frac{1}{70\sqrt{\pi n}} \right]^{1/5}$$

Kernel density is available using PROC KDE in SAS/Stat:

```
Proc KDE data=work.text;
Univar var1/out=outkde grid=0 gridu=10 bwm=1 method=SROT;
By population;
Run;
```

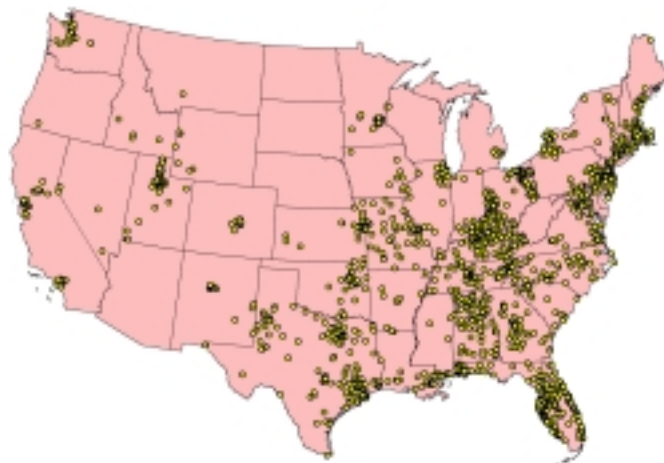
where method=... specifies the method to be used for optimizing the bandwidth; bwm is a multiplier that can be used to modify the optimal bandwidth, and gridl(u) specifies lower and upper bounds for investigating the kernel density estimator. A by statement can be used to investigate differences in sub-populations. Two examples combining geographic information with kernel density will be presented here.

Example 1.

The first example deals with a database containing approximately 2000 patient records with up to 5-years of follow up for patients treated for melanoma. A measure of UV exposure was found using geographic information to investigate the relationship of exposure to melanoma staging and patient laboratory data (as well as to examine patient quality-of-life). The UV-count is based on a regression equation from measurements at UV meters at 10 locations in the U.S. in 1974: UV count = 3220000 - 49613.9 * (latitude) + 104.3 * (altitude). Once computed, the UV count was scaled to a value between 1 and 2. (Fears et al. 1977): <http://aje.oupjournals.org/cgi/content/abstract/105/5/420>. The locations for the melanoma patients are given in Figure 4. Note that the patients are concentrated in specific geographic locations. This has more to do with the enrollment centers than it does with clusters of melanoma at those locations.

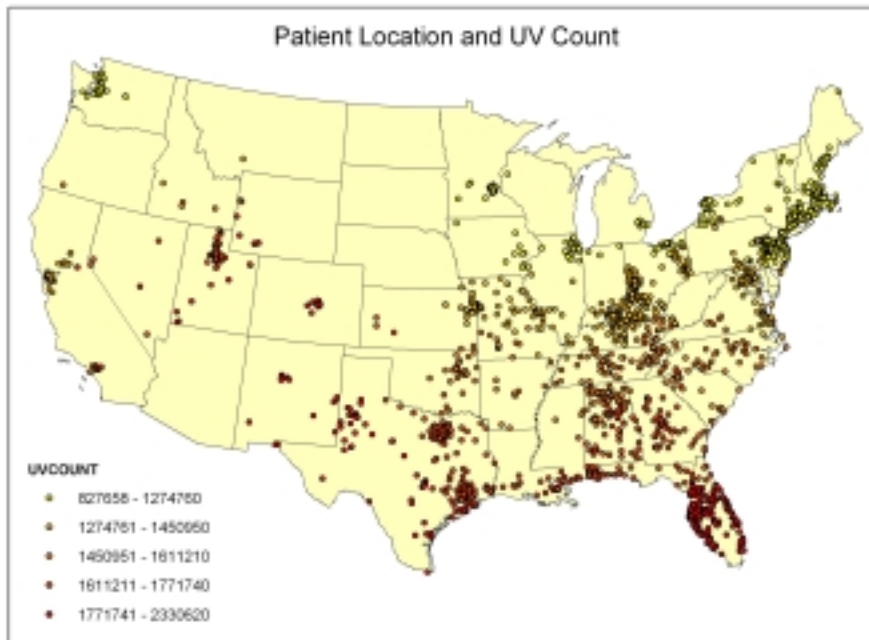
Figure 4

Locations of Zip Codes in Melanoma Database



It is possible to examine the relationship of type of melanoma to UV exposure. Figure 3 gives a map of the level of exposure for each patient using Fears et. al (1977) UV Count.

Figure 4. UV Index for Each Patient in the Database



As should be clear, the level of exposure decreases in relationship to the distance north of the equator, with some differing levels in mountainous areas. It is not possible to examine the number of cases in relationship to geographic location since the melanoma cases are strongly related to the cancer centers that enroll patients in a clinical trial, although it does give an indication of access of patients to clinical trials.

However, it is possible to examine the nature of the melanoma-its location, patient demographics, and so on. A general linear model was used to compare the relationship of gender, race, and factors defining the severity of melanoma to the level of UV exposure in the geography. However, because the number of melanoma cases is almost 3000, factors will be statistically significant because the number of degrees of freedom is so high (summarized in Table 1).

Table 1. Results of General Linear Model Analysis

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|------------------------|----|--------------|--------------|---------|--------|
| Race | 2 | 43980932077 | 21990466039 | 0.36 | 0.7012 |
| Gender | 1 | 74468124701 | 74468124701 | 1.20 | 0.2730 |
| Primary Site | 2 | 37589557856 | 18794778928 | 0.30 | 0.7381 |
| Vascular Invasion | 5 | 2.3969974E12 | 479399489476 | 7.74 | <.0001 |
| Vertical Growth | 2 | 987904818006 | 493952409003 | 7.97 | 0.0004 |
| Ulceration Present | 2 | 11055625074 | 5527812537 | 0.09 | 0.9146 |
| Evidence Of Regression | 2 | 750813024517 | 375406512258 | 6.06 | 0.0024 |
| Histologic | 6 | 900417836582 | 150069639430 | 2.42 | 0.0245 |
| Cutaneous | 1 | 13603528339 | 13603528339 | 0.22 | 0.6394 |
| BreslowThickness | 1 | 61470109305 | 61470109305 | 0.99 | 0.3193 |
| ClarkLevel | 1 | 196428568493 | 196428568493 | 3.17 | 0.0751 |

Therefore, the statistical significance needs to be validated. One way of validating the results is to examine kernel density estimates comparing factors. Kernel density also provides a method to relate a geographic measure to a patient characteristic. Figure 6 compares UV index (ie exposure) by gender. Note that the curves are almost identical. The similarity in the curves indicates that the difference by gender is not important, while also not statistically significant. Race is also not statistically significant. The factors that are significant are Vascular invasion, Vertical growth, Evidence of regression, and histologic. Clark level is marginally statistically significant. These graphs were automatically generated using Enterprise Guide.

Figure 6. Comparison of UV Exposure by Gender

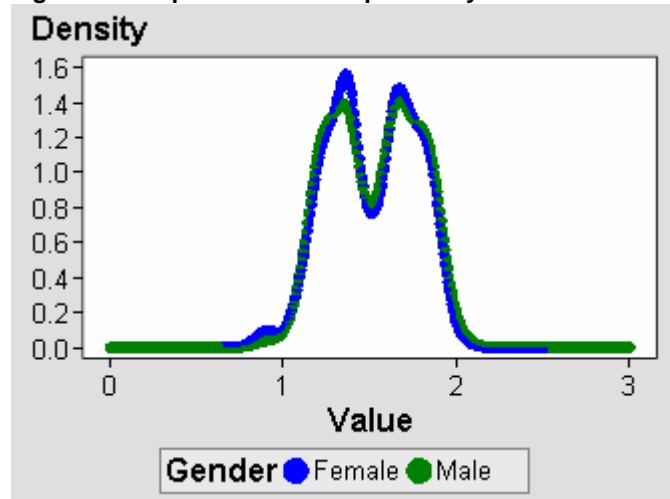


Figure 7 examines the relationship of UV index to the location of the melanoma: on the head and neck, on extremities (ie arms and legs), or on the trunk. There is a slight difference in the curves. The probability of a large index (close to 2) is higher for extremities and trunk locations compared to the head and neck. There is an interesting bimodality in the extremity category. Although it is permissible to speculate a little as to why (higher exposure to UV light is related to wearing less clothing, exposing more of the trunk and extremities), the speculation must be validated with additional data. However, the slight shift is not enough to yield statistical significance.

Figure 7. Comparison of UV Exposure by Melanoma Location

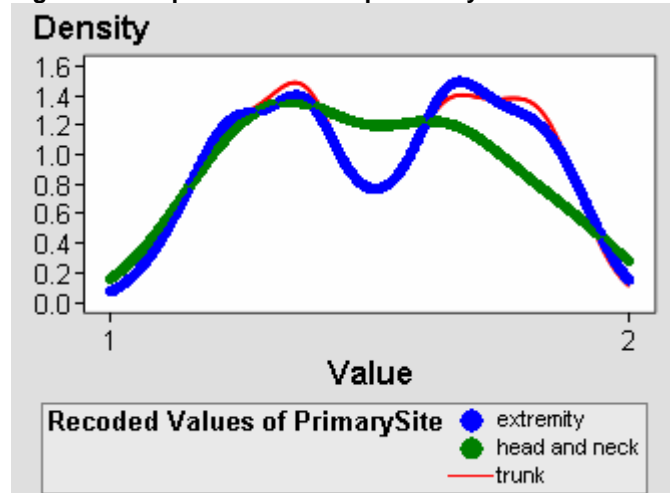


Figure 8 compares UV index by race. There is a problem with the graph. African Americans and Others occur in very small numbers compared to Caucasians. When using kernel density estimation for data visualization, the optimal bandwidth choice remains the same for each segment of the population. However, the bandwidth is a function of the sample size, and it increases in value as the sample size increases. Because there is such an extreme difference in the sample sizes for the three races, the optimal bandwidth is not optimal for all three groups. A second kernel density estimator was used decreasing the bandwidth to 0.25 of the optimal (Figure 9).

It is noted that the estimator for Caucasian is now too jagged, and the bandwidth is too small. On the other hand, the other two estimators are much more reasonable; the first estimates are over-smoothed. What is optimal then, is to use the estimator for Caucasian from Figure 8, and the other two estimators in Figure 9 to get a meaningful comparison of the graphs. Once this is done, it becomes clear that Caucasians and African Americans have a higher probability of a higher UV index compared to the Other category (Figure 10).

Figure 8. Comparison of UV Exposure by Race

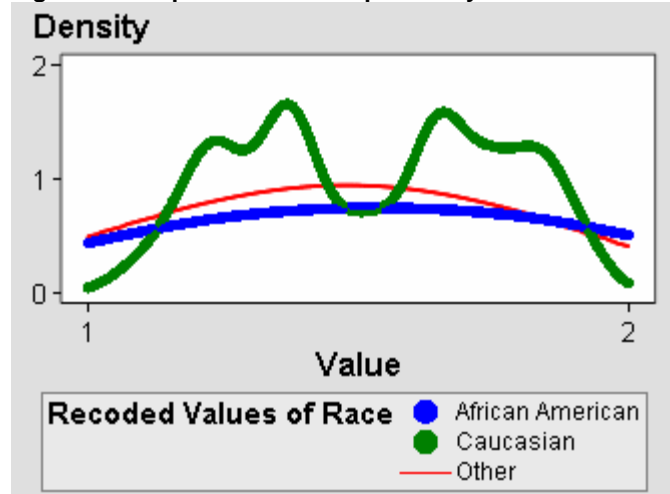


Figure 9. Reduction in Bandwidth

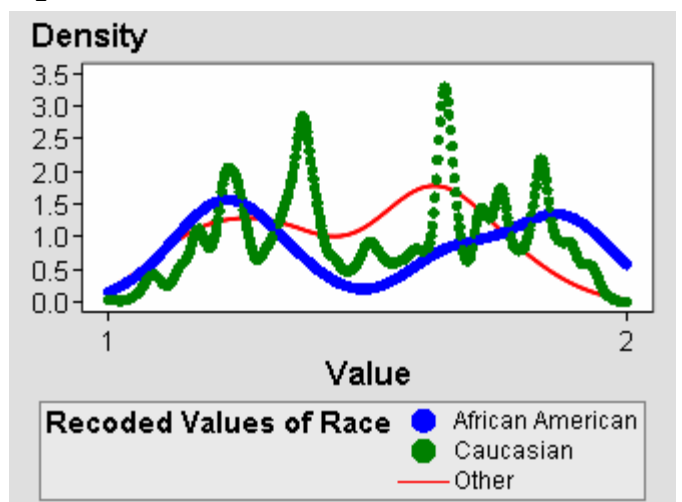
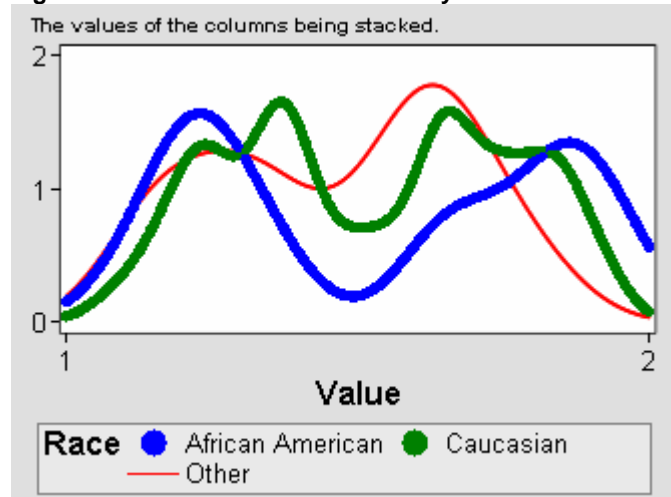


Figure 10. Combination Kernel Density for Race



Figures 11 and 12 examine the relationship of measurements of the severity of the melanoma compared to the UV index. Figure 11 examines the extent to which the melanoma has invaded blood vessels. Figure 12 examines whether there has been some vertical growth in the melanoma. In both cases, the more severe melanoma has a higher probability of a higher UV index. These two variables were highly statistically significant in the linear model, so the rightward shift is sufficient, as compared to the shift for primary site (Figure 7).

Figure 11. Comparison of UV Exposure by Level of Existence of Vertical Growth

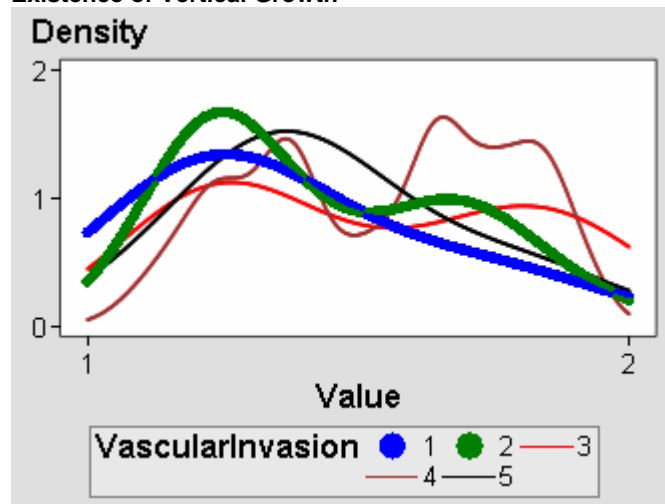
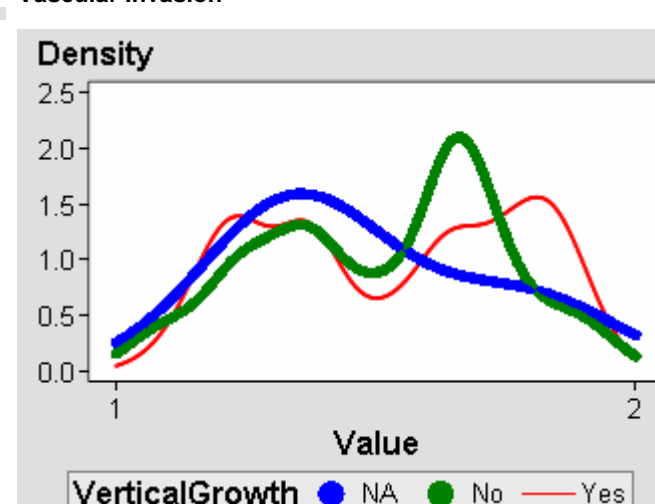


Figure 12. Comparison of UV Exposure by Vascular Invasion



In ArcGIS, kernel density estimation is used to create a continuous density surface using a set of input features, in this case points representing patient locations. In SAS, this is comparable to a bivariate density, where one of the variables is defined as distance. The bandwidth ('search radius' in ArcGIS terminology) is defined as a unit of distance. The size of the bandwidth determines the smoothness of the contour. In ArcGIS, a grid is overlaid over the points and the value of each grid cell is determined by examining and weighting all points within the specified search radius, or bandwidth. Figure 13 shows the resulting surface using a small bandwidth, using the value in the vascular invasion field. In this example, virtually every point in the data forms a contour. For a larger bandwidth, the contour plot is smoothed (Figure 14). Because data in ArcGIS are inherently geographic, ArcGIS shows the geographic location and provides map elements such as a legend.

Figure 13.

ArcGIS Kernel Density Function
Louisville (KY) MSA Patients
Search Radius (Bandwidth) = 0.014 ft.
Variable = Vascular Invasion

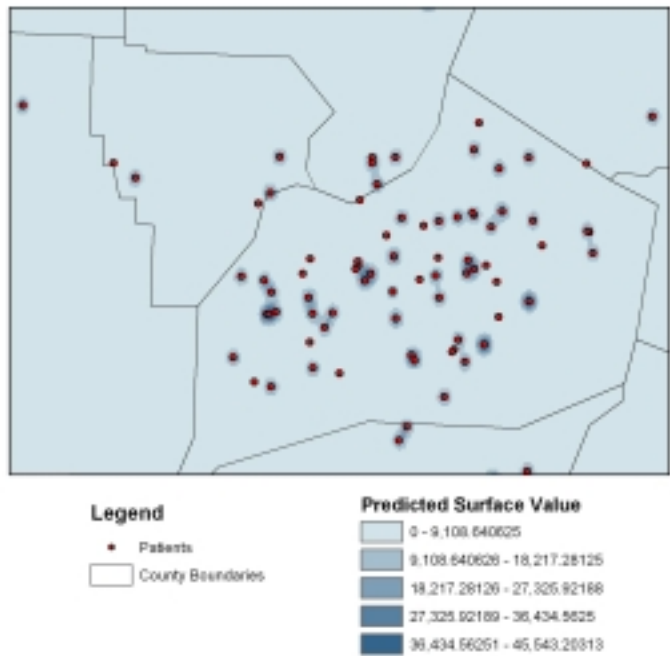
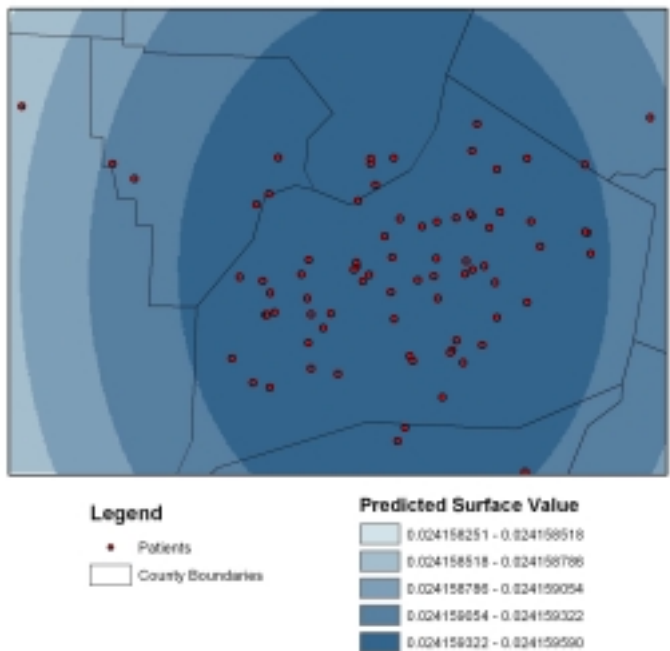


Figure 14

ArcGIS Kernel Density Function
Louisville (KY) MSA Patients
Search Radius (Bandwidth) = 100 ft.
Variable = Vascular Invasion



In SAS (version 9.1), PROC KDE can provide contour plot that is similar in concept but does not show the specific geographic location of input features, in this case, patient locations. The following code was used to produce the contour plots in Figure 15-17. Reducing the bandwidth by modifying bwm=.5 gives a much narrower contour plot (Figure 16)

```
proc kde data=sasuser.sunbeltwithuvcount;
bivar dist vascularinvasion/out=sasuser.kdebivariate3 bwm=2
plots=contour ;
where vascularinvasion ge 3;
run;
```

Figure 15. Contour Plot of Bivariate Kernel Density Estimation

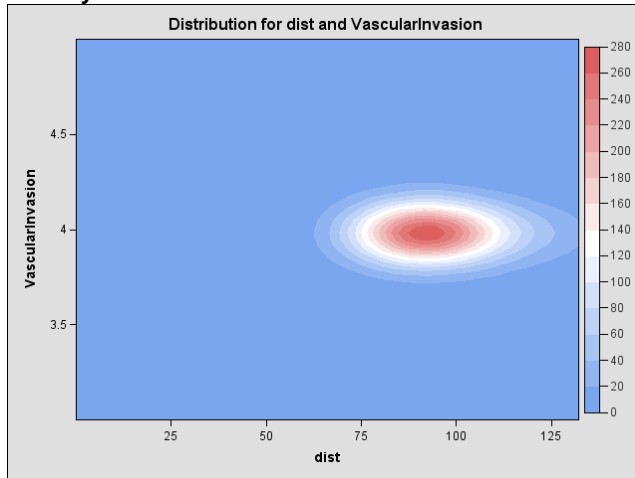
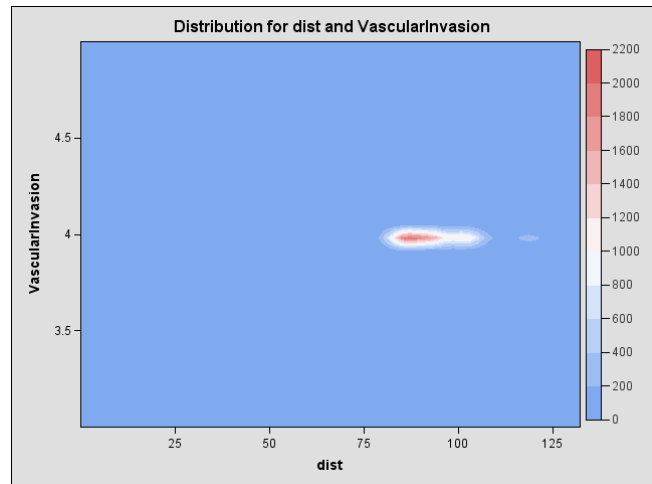
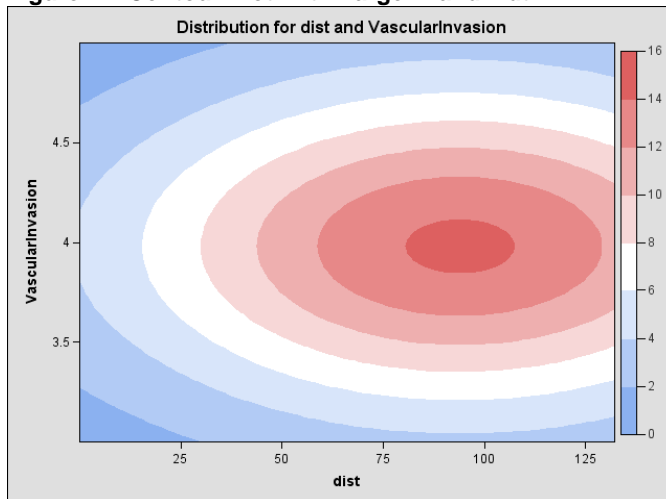


Figure 16. Kernel With Smaller Bandwidth



while increasing the bandwidth by bwm=5 yields the result in Figure 17

Figure 17. Contour Plot With Larger Bandwidth



Although the contour plot provided with kernel density estimation is readily available through PROC KDE, the relationship to geographic location is lost. Also, any sense of directional bias is lost. While the density estimates generated in Figures 13 and 14 are close to symmetrical, there is a slight North/South elongation. Were that more pronounced, estimations of density perpendicular to the major axis would differ considerably from those following the trend of the major axis. Directional bias would then be evident in ArcGIS but not in SAS.

Finally, SAS estimates do not account for edge effects. Neither does ArcGIS mathematically, but study area edge effects are at least subject to visual interpretation.

Example 2.

Consider the population of approximately 1000 patients were recruited to a randomized trial of lung cancer screening. To be eligible, patients needed a minimum of 40 pack years (1 pack per day for 1 year=1 pack year), and reduced pulmonary function as identified by spirometer testing. A questionnaire was used for any patient applying for the study. The questionnaire asked patients to self-report on lifetime and current smoking habits, and also on exposure to specific toxic chemicals. Consistently, African Americans reported lower numbers of cigarettes smoked both currently and in the past. Therefore, based upon this self-reporting, many African Americans were not eligible for the study.

As an additional criterion for the study, patients were not eligible for enrollment if a baseline x-ray showed a suspicious lesion that was diagnosed as positive after biopsy. In other words, all of the eligible patients had to be cancer-free for enrollment. Therefore, this was not a study to examine the relationship between environment hazards and disease incidence. It was not to validate exposure levels. The purpose of the study was to examine the relationship between self-reporting and actual, current exposure to determine whether there was a consistent pattern of under-reporting.

African Americans represented 11% of the population applying for the study but only 3.5% of the eligible population. The remaining 6.5% were ineligible because they did not meet the inclusion/exclusion criteria. It was hypothesized that inclusion/exclusion criteria should account for patterns in patient reporting and outcomes across race and ethnicity. Otherwise, clinical trials, particularly those that rely on self-reporting, will continue to have reduced numbers of minority populations.

To examine the issue of self-reporting of smoking habits, additional data were examined since African Americans also self-reported exposure to toxic chemicals in lower proportion compared to Caucasians (Table 2).

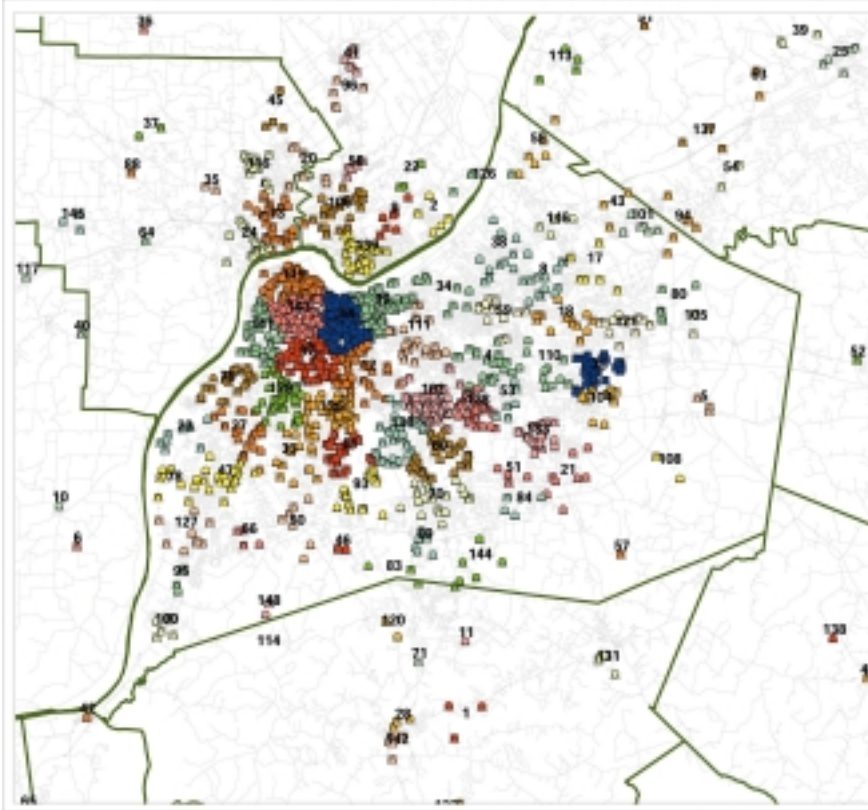
Table 2. Self-Reporting Exposure

| Exposure | African American | Caucasian | Not Known African American | Not Known Caucasian |
|-----------------------|-------------------------|------------------|-----------------------------------|----------------------------|
| Asbestos | 11 (13.2%) | 195 (16.8%) | 26 (31.3%) | 257 (22.1%) |
| Radon | 4 (4.9%) | 50 (4.3%) | 19 (23.5%) | 239 (20.8%) |
| Arsenic | 2 (2.5%) | 32 (2.8%) | 20 (25.0%) | 184 (16.1%) |
| Beryllium | 0 (0%) | 32 (2.8%) | 21 (26.2%) | 190 (16.6%) |
| Metals | 3 (3.7%) | 65 (5.7%) | 23 (28.7%) | 217 (18.9%) |
| Coal | 1 (1.2%) | 63 (5.5%) | 21 (26.2%) | 159 (13.8%) |
| Mustard Gas | 1 (1.2%) | 36 (3.1%) | 19 (23.7%) | 185 (16.2%) |
| Vinyl Chloride | 1 (1.2%) | 51 (4.5%) | 22 (27.5%) | 252 (22.1%) |

Note that 31% of African Americans do not know whether they were exposed to asbestos compared to 22% of Caucasians. Similarly, Caucasians report almost five times as much exposure to coal compared to African Americans. Therefore, it appears that under-reporting in the African American Community is general in terms of exposure to environmental hazards, including cigarettes. Since most models that predict risk depend upon accurate patient information, African Americans will generally be perceived as having lower risk than they actually have.

Unlike cigarette smoking, exposure to toxic substances can be examined more carefully. Using information generated by the EPA, it is possible to investigate the proximity of African Americans to toxic hazards. Figure 18 gives a map of the Louisville, Kentucky area with all toxic sites labeled. As shown in the map, toxic sites are concentrated in the West End of Louisville.

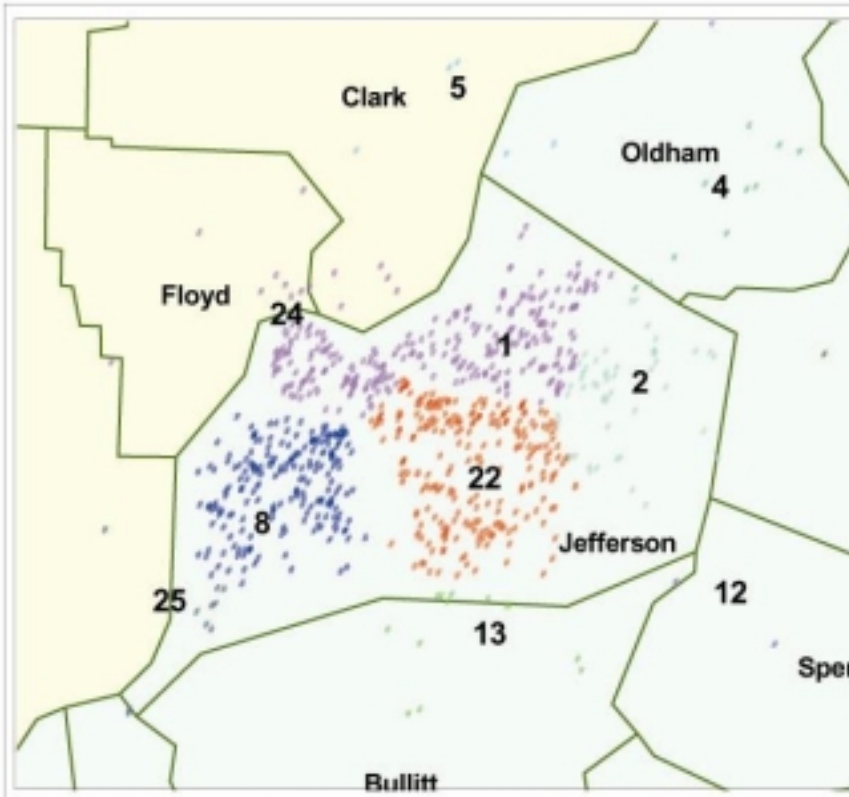
Figure 18. Toxic Sites in Louisville, Kentucky



Patient addresses were layered onto the map to define a measure of proximity to toxic release (Figure 18). The map was divided into equally spaced patient clusters using PROC FASTCLUS (Figure 19).

The X-Y coordinates were used to define the patient clusters. Because distance data were the only variables used, the clusters had relatively equally sized radii with seeds that were randomly identified using K-means clustering. The purpose of the clusters was to find natural geographic breaks in the patient base. Although distance values were used, the results represented non-spatial clustering. It did simplify the problem of exploring the relationship between actual exposure compared to the self-reporting of exposure in the identified clusters.

Figure 19. The Geographic Area Divided Using PROC FASTCLUS



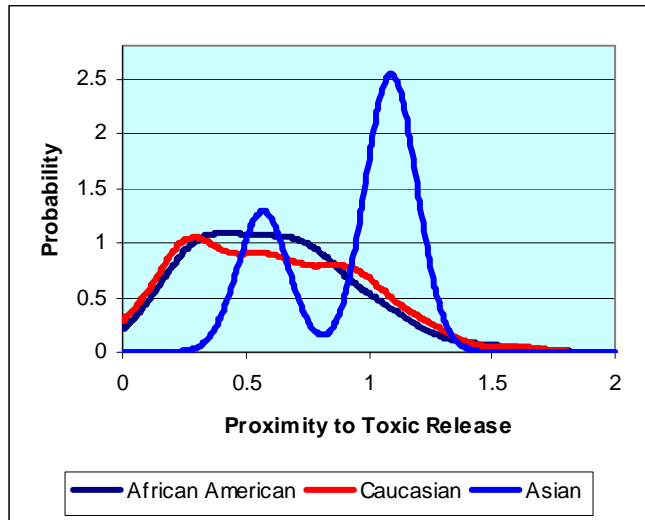
Patients in cluster 24 have the highest exposure to toxic release, and should report the highest level of exposure. Yet, restricting the data to cluster 24 indicates that African Americans still under-report exposure (Table 3).

Table 3. Reporting of Exposure

| Exposure | Caucasian | African American | P-Value |
|----------------|-----------|------------------|---------|
| Asbestos | 23.13 | 11.32 | 0.1339 |
| Radon | 3.75 | 3.77 | 0.343 |
| Arsenic | 1.88 | 1.89 | 0.3858 |
| Beryllium | 3.13 | 0.00 | 0.1135 |
| Metals | 8.13 | 0.00 | 0.0228 |
| Coal | 5.00 | 0.00 | 0.019 |
| Mustard Gas | 2.50 | 0.00 | 0.1297 |
| Vinyl Chloride | 5.00 | 0.00 | 0.0494 |

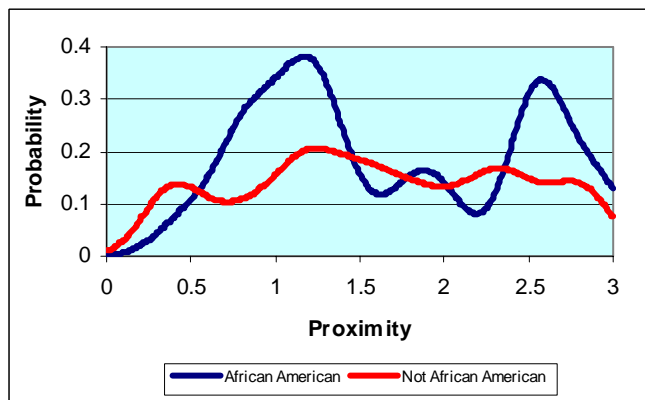
Note that although African Americans live in close vicinity to toxic substances, they self-report lower exposure; they tend to report lower levels of exposure. Some of that reporting of lower exposure is statistically significant.

Figure 20. Kernel Density of Proximity to Toxic Site by Race



Proximity is a geographic measure of the distance of one of the subjects to the nearest toxic site. Generally, African Americans have similar proximity to toxic release compared to Caucasians, with Asians more distant from the hazards. The difference is not statistically significant primarily because of the small number of Asians compared to the other two groups (Figure 20).

Figure 21. Kernel Density of Proximity Restricted to Cluster 24



However, in the area identified as cluster 24, the result is significant ($p < 0.0001$, Figure 21). Proximity by African Americans occurs primarily at a distance of one mile, and then again at a distance of 2.5 miles compared to Caucasians with more general variability.

Similarly, patient self-reporting of current and lifetime smoking were examined by identified geographic cluster (Figure 17). Note that cluster 24 has the highest reported proportion of current smokers with the lowest reported average lifetime smoking.

CONCLUSION

The SAS Bridge to ESRI greatly enhances the convenience of use when analyzing geographic data using SAS statistics procedures. ArcGIS is designed for providing spatial content while SAS is designed for mostly aspatial statistical analysis. Rather than trying to turn ArcGIS into a statistical program and turning SAS into a spatial engine, we need to focus on the ability to pass data between the two with the help of the SAS Bridge to ESRI, allowing each package to perform the tasks for which they were designed.

REFERENCES

1. K.Cromley E, McLafferty SL. *GIS and Public Health*. New York: Guilford Publications; 2002.
2. Fotheringham AS, Brunson C, Charlton M. *Quantitative Geography: Perspectives on Spatial Data Analysis*. Thousand Oaks, CA: Sage Publications; 2000.
3. Silverman B. *Density estimation for statistics and data analysis*. Boca Raton, FL: CRC Press; 1986.
4. Devesa SS, Grauman DJ, Blot WJ, Pennello G, Hoover RN, Fraumeni JF Jr. *Atlas of cancer mortality in the United States, 1950-94*. Washington, DC: US Govt Print Off; 1999 [NIH Publ No. (NIH) 99-4564].

APPENDIX

The syllabus for the 6-hour sequence of graduate courses is listed below.

Course Description: This is a two-course sequence in data mining and exploratory spatial data analyses using SAS and ArcGIS software. The sequence will be offered for 6 credit hours during the University of Louisville's Summer

Term III. The course will run Monday-Friday July 6-August 6, 9:00-am-4:00 pm (with a one-hour lunch break). After the introductory period, lecture materials will be presented in the morning; afternoons will be reserved for assignments and applications development. Each student will be required to complete a research project with a poster presentation. Students will be evaluated based upon class participation, homework assignments, and the final project presentation.

Preliminary Schedule:

| Dates | Topics | Applications & Development |
|-----------|--|--|
| 7/6-7/8 | Introduction to ArcGIS | <ul style="list-style-type: none"> • Age standardization |
| 7/9-7/13 | Introduction to SAS and statistics <ul style="list-style-type: none"> • Enterprise Guide & SAS Bridge to ESRI • General linear models • Mixed models • Cluster analysis • Kernel density estimation | <ul style="list-style-type: none"> • Choropleth mapping of disease rates • Areal interpolation • Dasymetric mapping • ArcGIS interpolation methods • Trend surface analysis |
| 7/14-7/15 | Concepts in medical geography & epidemiology | <ul style="list-style-type: none"> • Kernel density estimation • Spatial autocorrelation • Spatial SCAN statistics |
| 7/16 | Spatial interpolation | <ul style="list-style-type: none"> • Deprivation indices |
| 7/19-7/20 | Proximity | <ul style="list-style-type: none"> • Lyme disease modeling |
| | Kernel density estimation | <ul style="list-style-type: none"> • Data mining |
| 7/21-7/22 | Clustering of health events <ul style="list-style-type: none"> • Spatial autocorrelation • Spatial clustering (point & area methods) • Spatial SCAN statistic | |
| 7/23-7/26 | Indices of health disparities | |
| 7/27-7/28 | Indices of health disparities | |
| 7/29-7/30 | SAS data mining techniques | |
| 8/2-8/6 | Final projects and presentations | |

ACKNOWLEDGMENTS

The authors wish to acknowledge support from the National Science Foundation grant # 0327581, Data Mining and Geographic Information Systems (GIS) to Investigate Public Health Information.

CONTACT INFORMATION

(In case a reader wants to get in touch with you, please put your contact information at the end of the paper.)

Your comments and questions are valued and encouraged. Contact the author at:

Patricia Cerrito
 Department of Mathematics
 Jewish Hospital Center for Advanced Medicine
 University of Louisville
 Louisville, KY 40292
 502-560-8534/502-852-6826
 502-852-7132 (fax)
pcerrito@louisville.edu
www.math.louisville.edu

Carol Hanchette
 Department of Geosciences/Geography
 University of Louisville
 Louisville, KY 40292
 502-852-2699
carol.hanchette@louisville.edu
<http://www.louisville.edu/a-s/geog>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.