

Calculation of health disparity Indices Using Data Mining and the SAS Bridge to ESRI

Mussie Tesfamichael, University of Louisville, Louisville, KY

Abstract

Socioeconomic indices are strongly believed to be associated with the risk of disease. However, no consensus exists in the US regarding which area-based measure should be used to measure or monitor socio-economic inequalities in health. The purpose of this paper is to determine which area-based socioeconomic measures would be most appropriate for US public health surveillance to investigate in relationship to the incidence of disease. Geographic information systems (GIS) manage, analyze, and disseminate spatial data. Arc Map is used to display results of the analysis in a variety of formats, such as maps, reports and graphs. The SAS Bridge to ESRI is used to transfer the spatial information directly into SAS datasets. The specific example here is to examine the relationship between the rate of cancer and the various indices of social economic (SES) conditions for the study area consisting of Kentucky, Tennessee, North Carolina, Virginia and West Virginia. Linear models and cluster analysis in SAS are used in this problem to investigate the spatial data from Arc Map and to optimize the definition of a socioeconomic health index.

INTRODUCTION

The study data consist of 522 counties in five different states. We will use three different methods to predict the rate of cancer based on the different socioeconomic indices. The SAS Bridge to ESRI will transform the spatial data to SAS datasets so that inferential statistics in SAS can investigate how the different independent variables would predict the rate of cancer.

Although the application of linear models and cluster analysis is widely used in investigating data, the models have not been used regularly with Arc Map. To enhance the use of SAS with Arc Map, SAS has developed the Bridge to ESRI. In this paper we will demonstrate how the SAS Bridge to ESRI is used to transfer the spatial information directly into SAS datasets. The primary task of this process is to locate the predictors of cancer rates from different categorical variables.

The use of linear models enables us to perform analysis of variance when we have a continuous, dependent variable with independent classification variables, quantitative variables, or both. Besides the usual estimators and test statistics produced for a regression, a fit analysis can produce many diagnostic statistics. Collinearity diagnostics measure the strength of the linear relationship among explanatory variables and how this collinearity affects the stability of the estimates. Influence diagnostics measure how each individual observation contributes to determining the parameter estimates and the fitted values.

In matrix algebra notation, a linear model is written as $Y = X\beta + \varepsilon$ where y is the $n \times 1$ vector of responses, X is the $n \times p$ design matrix, β is the $p \times 1$ vector of unknown parameters, and ε is the $n \times 1$ vector of unknown errors.

Factor analysis selects which variables in the data set are explanatory variables. The main applications of factor analytic techniques are: (1) to *reduce* the number of variables and (2) to *detect structure* in the relationships between variables, that is, to *classify variables*. Therefore, factor analysis is applied as a data reduction or structure detection method.

With the SAS Bridge to ESRI, we can export data to a SAS data set and use SAS to perform any analysis that is needed. The SAS Bridge to ESRI adds the analytic intelligence of SAS to the easy-to-use mapping capabilities of Arc GIS. The result is a geographic information system unmatched in the ability to inform, persuade, and motivate. Also, we can join SAS data to Arc Map layers to uncover new relationships in the existing data to find answers and to solve problems [1].

BACKGROUND

Cervical cancer is the number one killer of women in many developing countries. More than 3900 women die in the United States each year from this disease. A woman who doesn't have screening on a regular basis significantly increases her chances of developing cervical cancer. Only 11% of women report that they do not have regular cervical cancer screenings [5]. Three different methods will be used to investigate the rate of cervical cancer in white females to demonstrate the use of the SAS Bridge. The data were obtained from www.cancer.gov. Data concerning geographical details for the study states of Kentucky, Tennessee, Virginia, West Virginia and North Carolina were obtained from www.census.gov. The main objective is to predict the rate of cancer based on the level of the predictor variables (Table 1)

Table 1. Predictor Variables for Cervical Cancer

Table Name	Variable Name	Description
Low Education	LOWED	Percentage of persons aged >=25 with less than High school education
High Education	HIGHED	Percentage of persons aged >=25 with at least 4 years of college.
High Occupation	HIGHOCC	Percentage of persons employed in predominantly working class
Low Occupation	LOWOCC	Low Occupation
Low Income	LOWINC	Percentage of households with an income <\$15,000.
High Income	HIGHINC	Percentage of population with an income >\$150,000.
Poverty	POVERTY	Below federally defined line: for example income below \$12,647 for a family of four
Crowding	CROWD	Household with more than one person per room
No Vehicle Available	NOCAR	Percentage of no car ownership
High House Value	HIGHVAL	Homes worth>=\$300,000

First, we download the Geographic files stGeo.uf3 from the www.census.gov and then download the files for the variable of interest of the study: Educational attainment (P037), Occupation (P050), Income (P052), Poverty (P087), Poverty Ratio (P088), Tenure by Person by Room (H022), Tenure by Vehicles (H044), Value of Housing (H084), are downloaded for each of the five states. For each county of the five states, Cervical Cancer data are downloaded as well.

The next step was to define each of the variables into quintiles (top 20%, next 20%, middle 20%, next lowest 20% and bottom 20%) in which the top is given a value of 1 and the bottom is given a value of 5, where a "1" represents the "best" one fifth of cases and a "5" the "worst" By default, SAS usually starts calculating percentiles at the low end, which is the reverse of the natural order of the quintiles. If a variable is reversed, that is, a high value doesn't equate to a "good" situation (for instance, a high percent of Poverty is not a good situation, the 1st percentile should be coded as 5, not 1.

General Linear Model

The first method we used in predicting the rate of cancer was to assign a level based on percentile for each predictor variable. We then classified people in public health databases by the socioeconomic characteristics of their residential neighborhood (here: Rate of Cancer by County.). The Index for the cancer rate was calculated from all the variables and level was assigned to the index based on the percentile level. The index level1 was used to predict the rate of cancer in white females across the states of study: Kentucky, Virginia, West Virginia, and North Carolina. The following SAS procedure was used for this task.

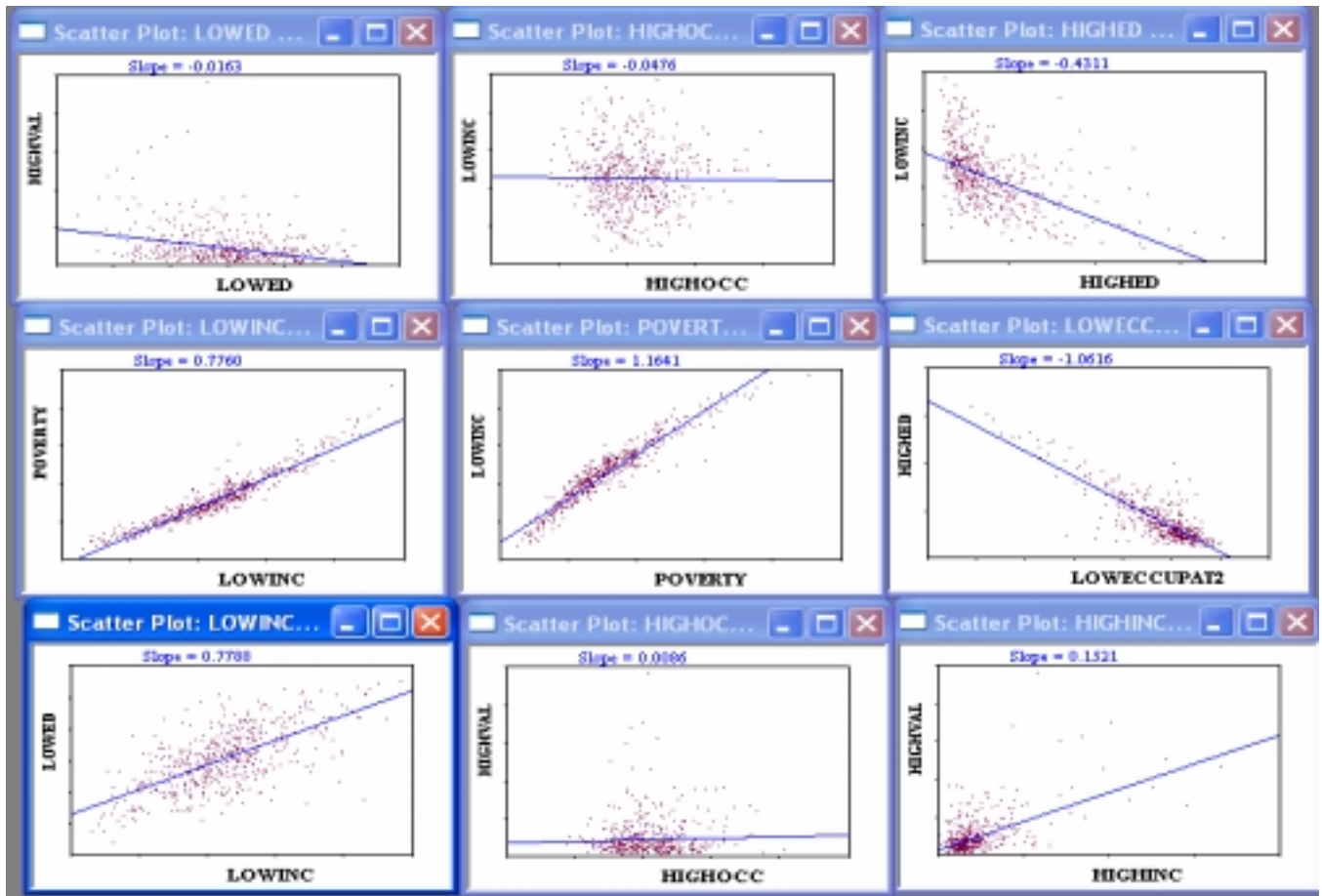
```
proc GLM data=SASUSER.SepIndexlevel1;
class indexlevel1;
model RATEWFLEVEL_NUM = indexlevel1 /solution;
output out=indexlevel1 p=_pred ;
run;
data sasuser.predRATEWFInd1;
set indexlevel1;
PredValue1=round(_pred);
Run;
```

Sas Output

R-Square		Coeff Var		P-value	
0.075029		45.6143		<0.0001	
		Param	SE	Pr > t	
Intercept		3.48	0.132	<.0001	
indlevel1 1		-1.06	0.180	<.0001	
indlevel1 2		-0.60	0.191	0.0018	
indlevel1 3		-0.18	0.190	0.3532	
indlevel1 4		-0.44	0.195	0.0232	
indlevel1 5		0.00	.	.	

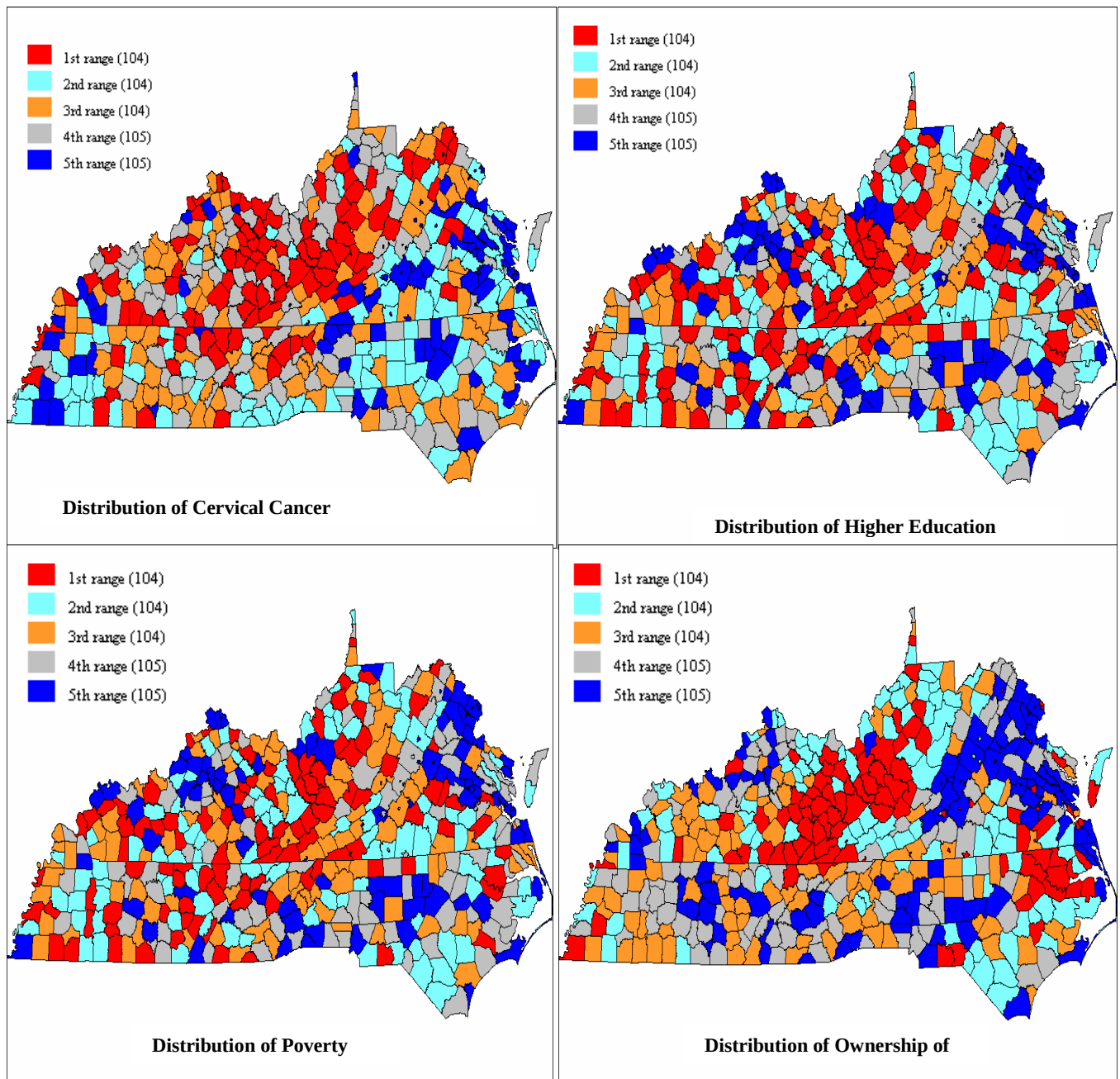
Result 1

R^2 for indexlevel1 was 0.075029 and $P < 0.0001$. This tells us that something is wrong. The problem is Multi-collinearity. The variables that produced the socioeconomic indexlevel1 were highly correlated, but this was not taken into consideration in calculating Indexlevel1. Even though the overall P (0.0001) value is very low, most of the individual P values are high. This would suggest that the model doesn't fit well, even though none of the predictor variables has a statistically significant impact on predicting the rate of cancer.



in a similar way.

Figure 2. Distribution of Cervical Cancer, Higher Education, Poverty and Ownership of Cars



These maps show that the eastern part of Kentucky and the northern part of Tennessee have very high rates of cancer and very low rates of higher education. In contrast, Virginia has a very high rate of education and lower rates of cancer.

Mapping the distribution of cervical cancer rates in the general white female population compared to cervical cancer rates for white females in poverty showed that eastern Kentucky and northern Tennessee have high rates of poverty, which is not the case in the eastern portion of Virginia.

II Factor Analysis

Since the GLM method of linear models didn't give a good fit of the data, a factor analysis was used. The factor analysis gave two different factors. The variables related to Economic resources are grouped together as Factor 1 and those that are related with Employment and Education are grouped in Factor 2. The SAS code gave the following result for the standardized scoring coefficients

SAS out put		
	Factor1	Factor2
POVERTYlev_num	0.30009	-0.12257
LowINCleve_num	0.26076	-0.04840
NOCARlevel_num	0.32712	-0.24129
HighINClev_num	0.09468	0.17643
HIGHVALlev_num	0.13587	-0.00770
HighEdleve_num	0.00512	0.29817
HighOcclev_num	-0.22325	0.38763
LowEdlevel_num	0.05249	0.22684
CROWDlevel_num	0.06763	-0.17714
LowOccleve_num	-0.01535	-0.26600

Now we construct an equation for index level2 by choosing the highest absolute value for each of the predictor variables:

$$\text{Factor1} = \text{POVERTYlev_num} * 0.30009 + \text{LowINCleve_num} * 0.26076 + \text{NOCARlevel_num} * 0.32712 + \text{HIGHVALlev_num} * 0.13587$$

$$\text{Factor2} = \text{HighINClev_num} * 0.17643 + \text{HighEdleve_num} * 0.29817 + \text{HighOcclev_num} * 0.38763 + \text{LowEdlevel_num} * 0.22684 - \text{CROWDlevel_num} * 0.17714 - \text{LowOccleve_num} * 0.26600$$

The standardized scoring coefficient gave a coefficient of 0.59762 for factor1 and factor2. Based on these results, the indexlevel2 of the rate of cancer is calculated. Factor analysis gave the following result:

SAS out put				
R-Square	Coeff	Var	P-value	
0.110648		44.727		<0.0001
Interc	Param	SE	Pr > t	
indlevel2 1	-1.41	0.185	<.0001	
indlevel2 2	-0.69	0.185	0.0002	
indlevel2 3	-0.52	0.185	0.0052	
indlevel2 4	-0.33	0.186	0.0791	
indlevel2 5	0.00	.	.	

Result 2

R^2 for indexlevel2 was 0.110648 and $P < 0.0001$. The overall P value is significant, and only one of the indexlevels is marginally significant (0.0791). The factor analysis, then, improves upon the first model. When cluster analysis is performed, the five index levels are clustered into three different classes.

Clusters	2	3	4	Total
1	44	49	12	105
2	22	74	8	104
3	18	65	22	105
4	16	57	31	104
5	4	69	31	104
Total	104	314	104	522

Table 2: Predicted Cervical Cancer classes by Indexlevel2 where only three classes are observed.

III Interaction effect

The previous two methods didn't give good results; as a consequence we are forced to seek another method. So we introduced an interaction effect on two groups where one can be called Low social Class variables and High social class variables. This gave a better result still. The following SAS code was used

```
proc glm data=SASUSER.Interaction;
  class HIGHVALLEV LOWEDLEVEL HIGHEDLEVE LOWOCCLEVE HIGHOCCLEV LOWINCLEVE
    HIGHINCLEV POVERTYLEV CROWDLEVEL NOCARLEVEL;
  model RATEWFlevel_num= HIGHVALlev*HIGHEDleve*HIGHOCClev*HIGHINClev
    LOWOCCleve*LOWINCleve*POVERTYlev*CROWDlevel*NOCARlevel*LOWEDlevel /solution;
  output out=sasuser.method3data p=_pred;
run;
data sasuser.allmethod;
set sasuser.method3data;
predvalue3=round(_pred);
run;
```

Result 3

The prediction table obtained by using the interaction effect gave five clusters as desired. As long as many interaction effects are included, the model is going to fit the data. But one must carefully consider the case, as more interaction is included that the model might over-fit the data.

Frequency Col Pct	1	2	3	4	5	Total
1	104 100	0 0	1 0.91	0 0	0 0	105
2	0 0	102 99.03	2 1.82	0 0	0 0	104
3	0 0	1 0.97	102 92.73	2 1.94	0 0	105
4	0 0	0 0	5 4.55	97 94.17	2 1.96	104
5	0 0	0 0	0 0	4 3.88	100 98.04	104
Total	104	103	110	103	102	522

Table 3. Prediction Table of the cancer rate based on the Interaction Method .The interaction method of predicting the rate of cancer provides almost a perfect clustering of the counties with very few misclassifications.

The local Moran is related to the interaction model since it controls the other variables (spatial regression). The local Moran test (Anselin 1995) detects the local spatial autocorrelation for the General linear model and Factor analysis. It can be used to identify local clusters (regions where adjacent areas have similar values) or spatial outliers (areas distinct from their neighbors). Local Moran can be used to investigate local spatial clusters and as a diagnostic for outliers with respect to the measure of global association (local instability). The Local Moran value for each observation gives an indication of the extent of significant spatial clustering of similar values around that observation. The Local Moran statistics are used to identify regions that differ significantly from those expected under the null hypothesis [3].

$I_{i,t} = m_{i,t} \sum_j w_{i,j} m_{j,t}$. The Local Moran statistic $I_{i,t}$ will be positive when values at neighboring locations are

similar, and negative if they are dissimilar. STIS (Space Time Information System) evaluates the significance of Local Moran statistic values with Monte Carlo randomizations, using conditional randomization.

$m_{i,t}$ is the z-score standardized dataset being tested for region i at time t.

$m_{j,t}$ is the z-score standardized dataset for region j at time t.

w_{ij} is a spatial weight set denoting the strength of connection between areas i and j.

GeoDa was used to investigate the spatial autocorrelation of the predictor variables. The resulting map shows the significant locations by type of association, and the significance map shows the locations in different shades of green,

depending on the degree of significance. The map consists of visualization explanation and exploration of interesting patterns in geographic data.

Figure 3. Moran scatter plot matrix and serial correlation for indexLevel1 and IndexLevel2

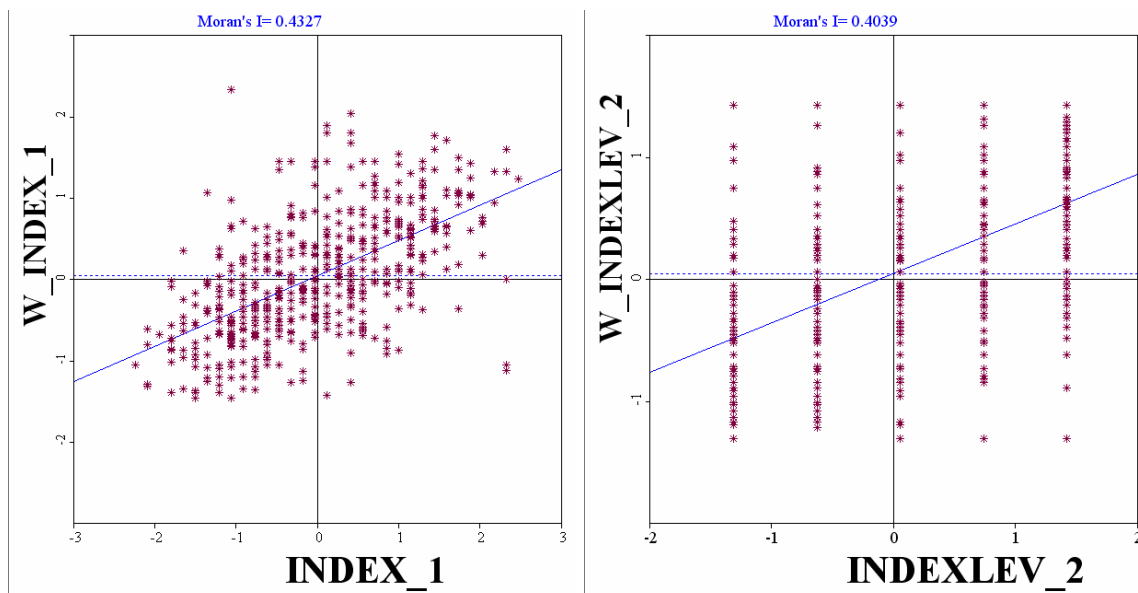
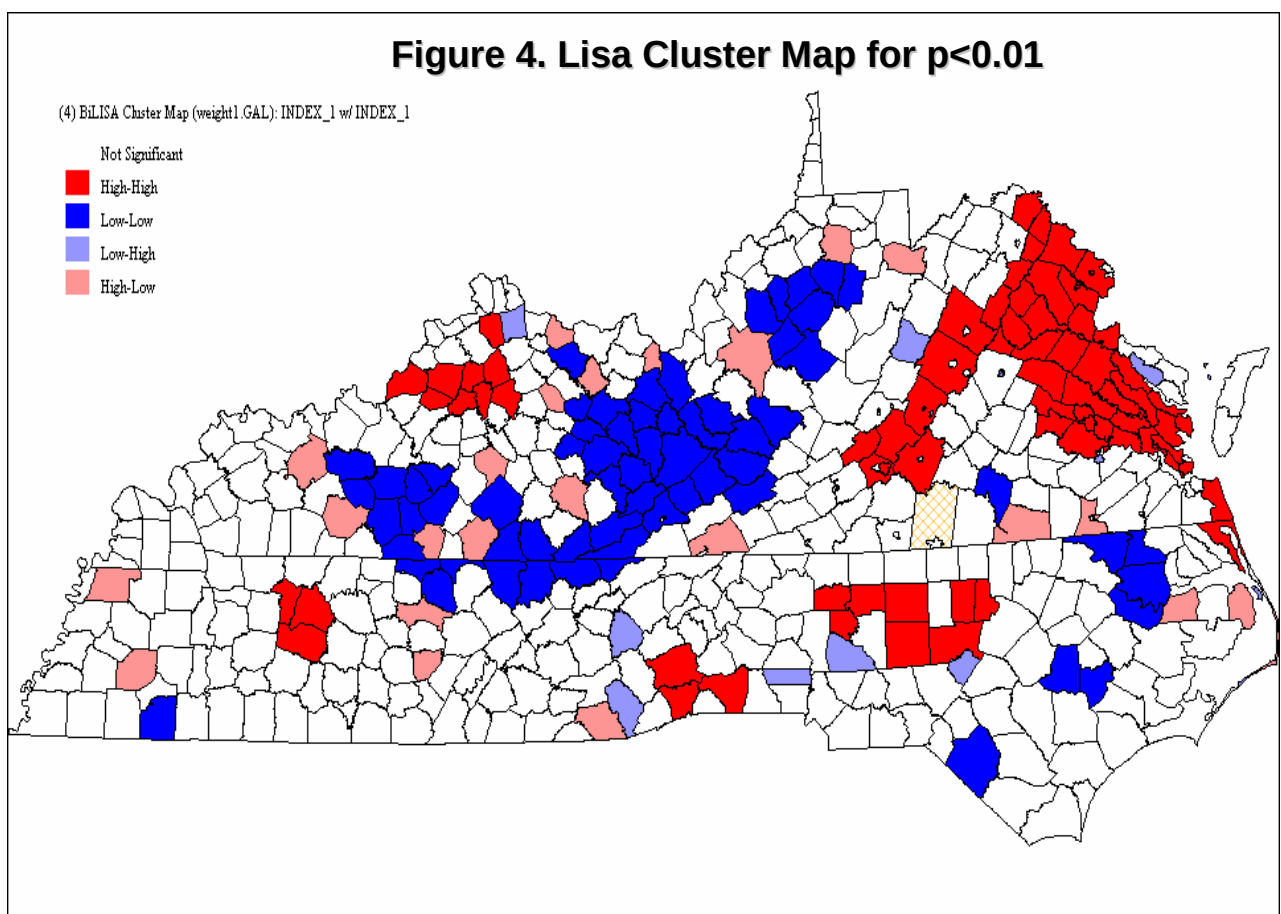
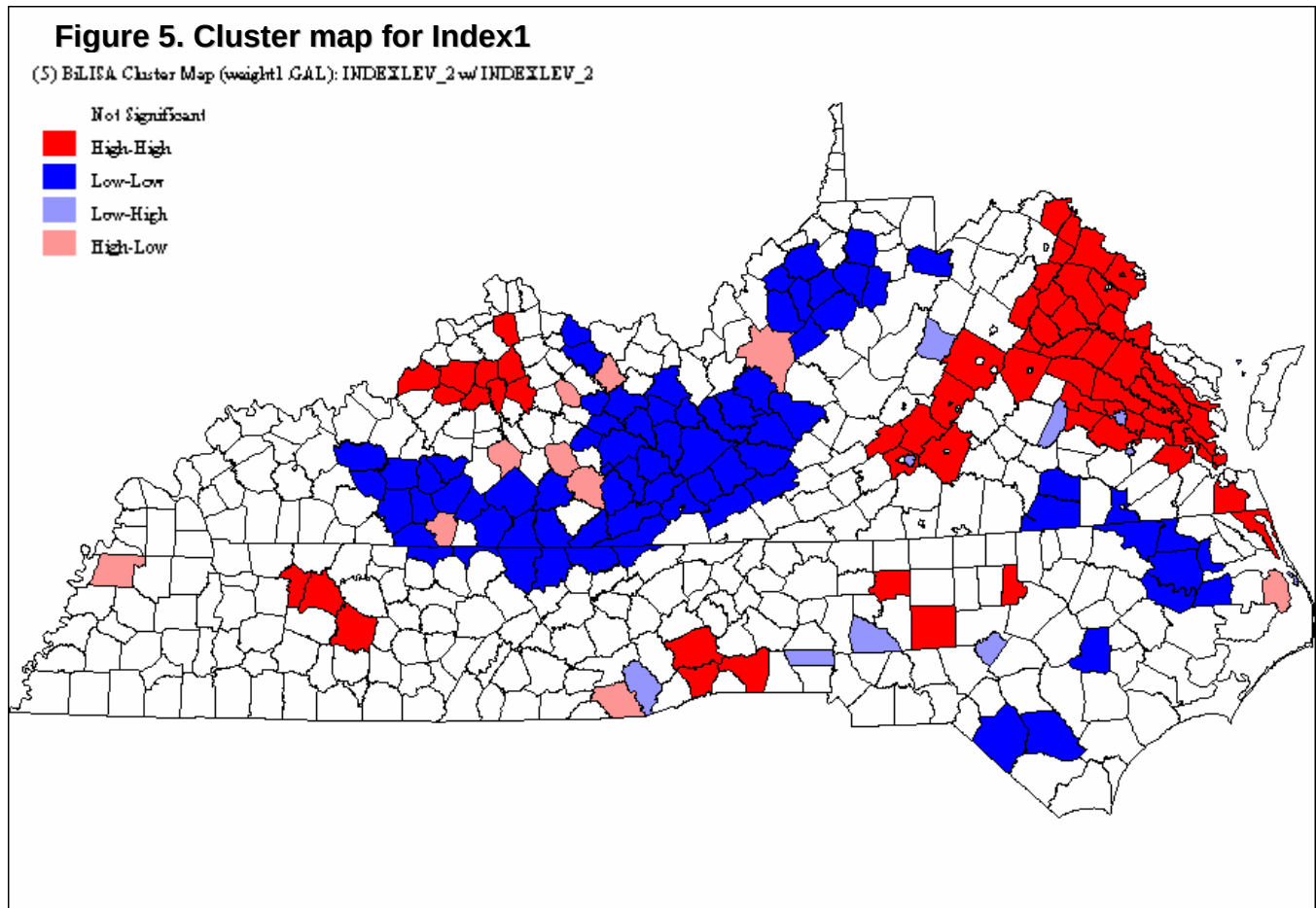


Figure 3 Accesses relationship between the variable value for unit of origin (x-axis) against the average of the values of its neighbors (y-axis).



High-High can be interpreted, as "I'm high and my neighbors are high." High-Low can be interpreted, as "I'm a high outlier among low neighbors", Low-Low can be interpreted "I'm low and my neighbors are low", and Low-High can be interpreted as "I'm a low outlier among high neighbors."

The map contains information only on those locations that have a significant Local Moran statistic. While every region in the dataset will be represented in the Moran Scatter plot, only those with Local Moran statistic p-values below 0.05 are colored red or blue on the example map above. Regions with non-significant Local Moran statistics are colored gray. [4]



CONCLUSION

No consensus exists in the US regarding which area-based measure should be used to measure or monitor socio-economic inequalities in health. The populations of the study states are highly affected by cancer. We were trying to find out if the risk of cancer is related with socio economic status. As we investigated, the General Linear Model and Factor analysis didn't produce a satisfactory prediction. So we continued our search method. The interaction method did well when compared to the other two methods, as it classified the rate of cancer into five different levels based on the socio economic indices. We conclude the interaction method gives a good prediction with a small improvement as a prediction of the rate of cancer. But one must carefully consider the case as more interaction is included and the model might over fit the data.

REFERENCES

- [1] ESRI and the ESRI globe logo are trademarks of Environmental Systems Research Institute, Inc.
- [2] Copyright Stat Soft, Inc, 1984-2003 STATISTICA is a trademark of Stat Soft, Inc.
- [3] **Reference:** Anselin, L. Local indicators of spatial association-LISA, 1995. *Geographical Analysis*, 27:93-115.
- [4] http://www.terraseer.com/products/stis/help/Statistics/LM/Results/Interpreting_univariate_Local_Moran_statistics.htm
- [5] National Cervical Cancer Coalition, <http://www.nccc-online.org/>.

CONTACT INFORMATION

Mussie Tesfamichael

Department of Mathematics

University of Louisville

Louisville, KY 40292

Work Phone: 502-852-7012, 502-298-8240

Fax: 502-852-7132

Email: matesf01@louisville.edu