Does Green Go with Gold?

Environmental assessment vs. economic development in the United States

—by Sumeeta Srinivasan

Introduction

Problem

How do the spatial distributions of environmental variables versus economic variables differ in the United States? This exercise examines economic and environmental indicators by state to determine whether there is evidence of spatial clustering of economic development using environmental assessment variables. Principal component analysis (PCA) is used to identify and remove redundancy (through correlation) in the data. The results of the PCA are then imported into ArcGIS® software to interpret them.

Location

United States

Time to complete the lab

Four hours

Prerequisites

- Basic understanding of principal component analysis
- Moderate familiarity with using ArcGIS 10
- Basic familiarity with the software package R

Keywords: environmental variables; economic variables; principal component analysis; Moran's I; local cluster analysis

Data used in this lab

The data consists of values for multiple variables for each of the 50 states in the United States. Some of the variables are economic indicators, and some are environmental indicators.

- USAStates: A shapefile showing all the states in the United States
  - Geographic coordinate units: Decimal degrees
  - Horizontal datum name: North American Datum of 1983
  - Ellipsoid name: Geodetic Reference System 80
- GREENGOLD_Table.csv: A comma-separated values table that includes all the variables of interest for comparing economic and environmental measures by state in the United States (Note that all the variables beginning with R are rank indicators for that attribute.)
- GREENGOLD_Table.xls: A Microsoft Excel table that includes descriptions of the various economic and environmental indicators

Student activity

One of the issues that is key to the assessment of the environment is the correlation between economic development and environmental awareness. In other words, is being "green" associated with wealth? To test this in the context of the United States, you will use statistical analysis to map various economic development and environmental indicators. Then you will carry out a statistical procedure called principal component analysis that will reduce the redundancy of these variables. You will then interpret what these newly derived attributes or components mean using the loadings[1] or coefficients of the economic and environmental indicators on these components. Lastly, you will map these PCA-derived components to see whether there is clustering in certain parts of the United States.

The steps that you will follow are

1. Examine the overall spatial patterns of economic development and environmental quality in the USA by mapping using ArcGIS.
2. Apply PCA to obtain uncorrelated variables that provide the same information as the original variables. The software program R will be used for the PCA.
3. Join these new variables to the shapefile and look for the spatial patterns of these newly derived measures.

[1] Loadings are the weights by which each standardized original variable for wealth and environmental measures should be multiplied to get the component scores that one gets after PCA.

Prepare your workspace

Data preparation, storage, and backup are basic and crucial when doing a geographic information system (GIS) project. It is good practice to store all your data within a single folder on your computer or storage device.

To begin, create a workspace to keep all data for this lab.

1 Create a folder for this lab exercise in your *C:\MyDocuments* folder. (For example, you could call it *EconEnvironAssess*).

2 Create a *Data* folder inside this folder.

Collect and process data

1 Download the data for this exercise to your *Data* folder.

2 Review the data before moving on to the analysis.

3 Launch ArcCatalog™.

4 Examine the shapefile and .csv file attributes (for all the US states and the GREENGOLD data, which includes all the environmental and economic data associated with the US states).

Note that the data variables are fully described in the Excel file.

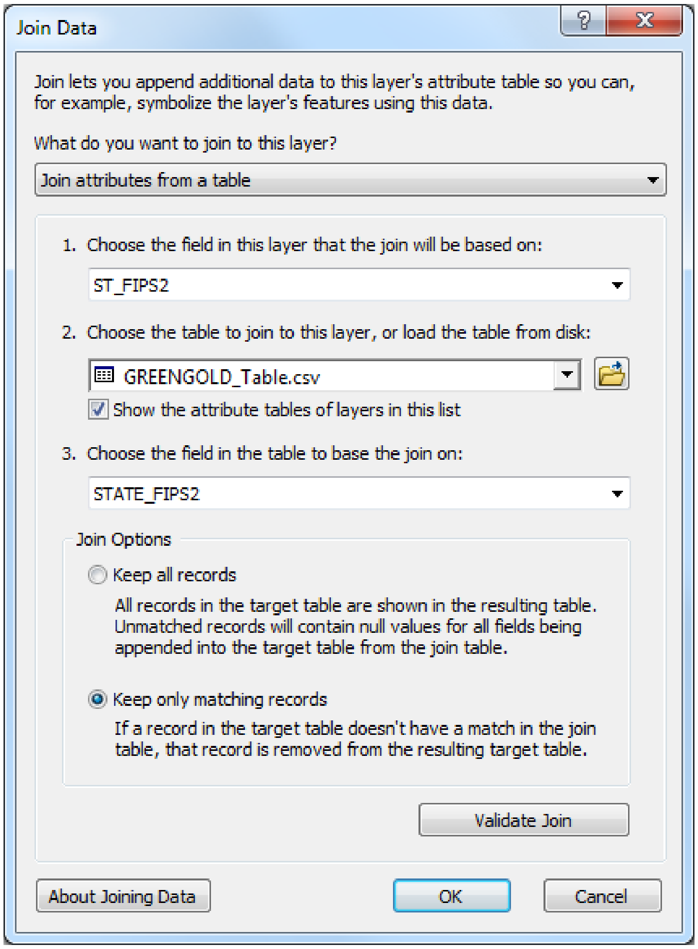
ASSESS ECONOMIC VS. ENVIRONMENTAL INDICATORS BY STATE

ANALYZE

1 Examine the attributes to see the spatial patterns of distribution.

2 Launch ArcMap™ and add the shapefile *USAStates* and the .csv file *GREENGOLD_Table.csv*.

3 In the table of contents, right-click the shapefile and then click *Joins and Relates » Join*. See the screen capture of the *Join Data* window to set up the join between the table and the shapefile as shown.



Join lets you append additional data to this layer's attribute table so you can, for example, symbolize the layer's features using this data.

What do you want to join to this layer?

Join attributes from a table

1. Choose the field in this layer that the join will be based on:

ST_FIPS2

2. Choose the table to join to this layer, or load the table from disk:

GREENGOLD_Table.csv

☑ Show the attribute tables of layers in this list

3. Choose the field in the table to base the join on:

STATE_FIPS2

Join Options

○ Keep all records

All records in the target table are shown in the resulting table. Unmatched records will contain null values for all fields being appended into the target table from the join table.

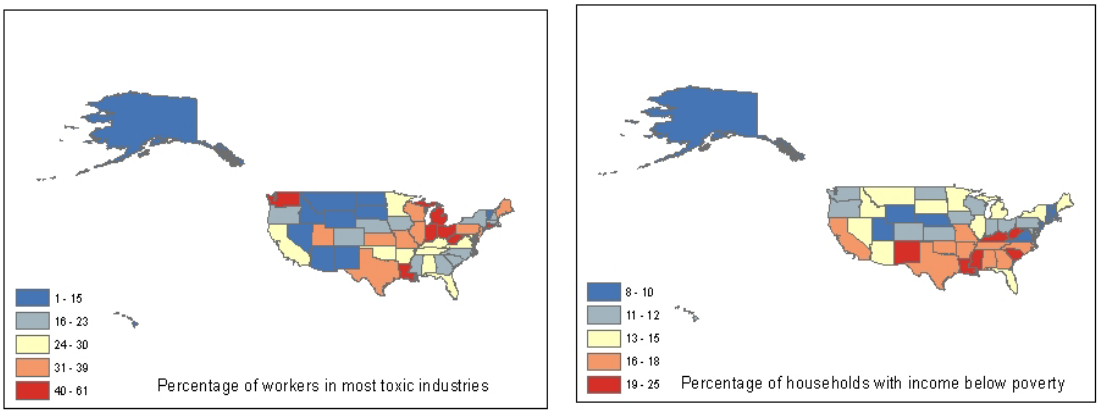
● Keep only matching records

If a record in the target table doesn't have a match in the join table, that record is removed from the resulting target table.

4 Right-click the shapefile and click *Properties » Symbology* and then *Quantities » Graduated colors*. Change the *Value* field to *WRKTOXIC* or *HHPOVERTY* to map some of the variables that you have from the joined table.

VISUALIZE

Map 1 shows the distribution of poverty by state versus the distribution of workers in the most toxic industries.


1 - 15
16 - 23
24 - 30
31 - 39
40 - 61
Percentage of workers in most toxic industries
8 - 10
11 - 12
13 - 15
16 - 18
19 - 25
Percentage of households with income below poverty

Map 1: Distribution of poverty vs. distribution of workers

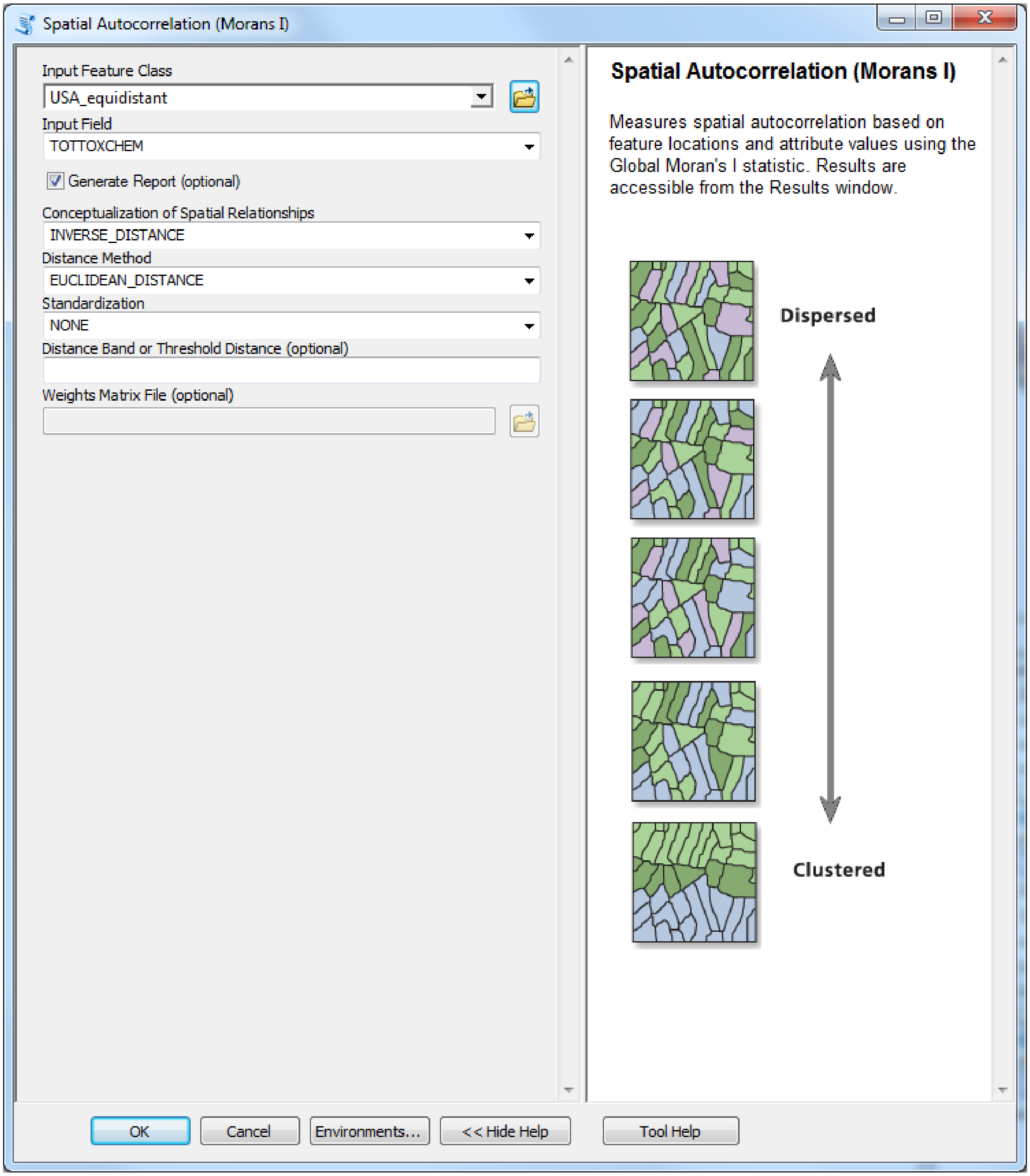
Question 1: *Do high percentages of workers in toxic industries and high poverty levels occur in the same states? Where are these variables spatially clustered?*

Question 2: *Choose some other economic development variables and environmental variables and then conjecture as to whether they are likely to be clustered spatially. Use maps to confirm or rebut your conjectures. How easy or difficult is it to confirm your hypotheses?*

ASSESS CLUSTERING OF ECONOMIC AND ENVIRONMENTAL DATA BY STATE

ANALYZE

1 You can assess the extent to which these variables may be clustered globally by computing the Moran's I statistic. In ArcToolbox™, expand *Spatial Statistics Tools » Analyzing Patterns*, double-click *Spatial Autocorrelation (Morans I)*, and select the variable you wish to assess for clustering as shown in the screen capture that follows. For example, the Moran's I is 0.09 and appears to be significantly clustered for total toxic chemical discharge. Try this for other variables of interest. Note the warning on the measures of distance. You may want to convert this dataset to a projection that is accurate for distance measures such as azimuthal equidistant projection (used by the US Geological Survey) or North America equidistant conic (available in ArcGIS).



Spatial Autocorrelation (Morans I)

Input Feature Class: USA_equidistant

Input Field: TOTTOXCHEM

☑ Generate Report (optional)

Conceptualization of Spatial Relationships: INVERSE_DISTANCE

Distance Method: EUCLIDEAN_DISTANCE

Standardization: NONE

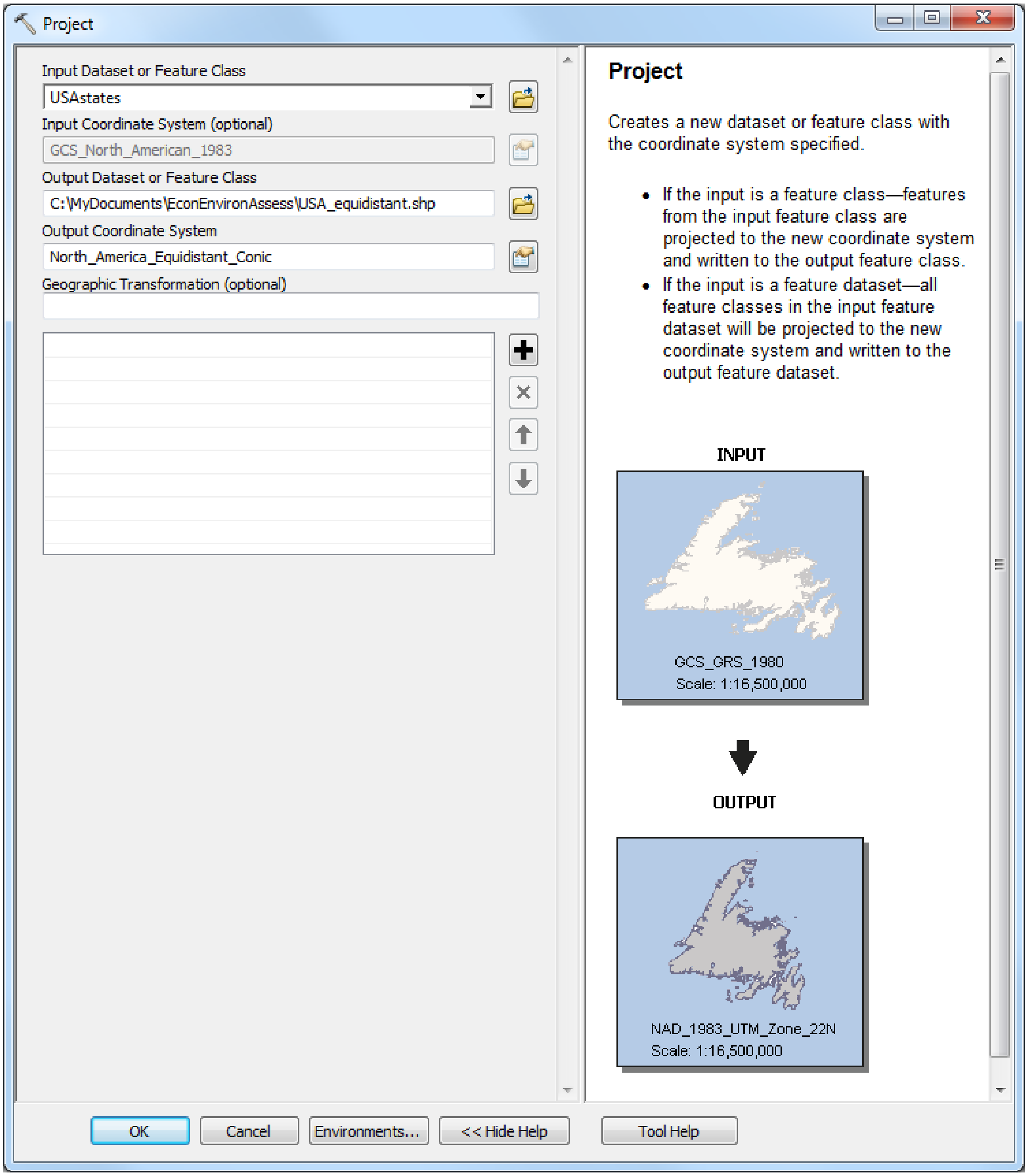
Distance Band or Threshold Distance (optional):

Weights Matrix File (optional):

Spatial Autocorrelation (Morans I)

Measures spatial autocorrelation based on feature locations and attribute values using the Global Moran's I statistic. Results are accessible from the Results window.

2 To convert the projection, use the *Projections and Transformations* toolset in the *Data Management* toolbox. Click *Feature » Project*. See the screen capture for details.



Project

Input Dataset or Feature Class: USAstates

Input Coordinate System (optional): GCS_North_American_1983

Output Dataset or Feature Class: C:\MyDocuments\EconEnvironAssess\USA_equidistant.shp

Output Coordinate System: North_America_Equidistant_Conic

Geographic Transformation (optional):

Project

Creates a new dataset or feature class with the coordinate system specified.

- If the input is a feature class—features from the input feature class are projected to the new coordinate system and written to the output feature class.
- If the input is a feature dataset—all feature classes in the input feature dataset will be projected to the new coordinate system and written to the output feature dataset.

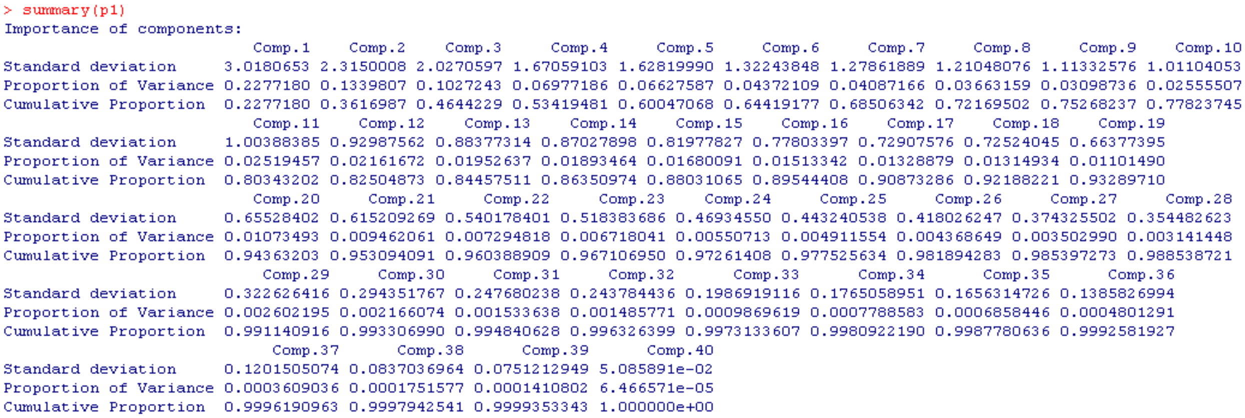
Question 3: *Report the Moran's I for several variables of interest.*

There are many variables in this dataset, and because they may be correlated, they may be providing the same information. The dataset would be easier to analyze if the variables were uncorrelated (also called orthogonal). PCA can be used to generate a (smaller) set of uncorrelated variables. Each original variable can be written as a linear combination of the uncorrelated variables. This helps simplify the information you get from the dozens of variables available to a few relevant characteristics that combine the variables that are correlated. Thus, you do not have to keep track of all the variables but only the combinations of variables that are most significant. You can do this in R, which is a free statistical package. Most statistical packages include an option for PCA. Be aware that the output may look slightly different from that for R even though the results are equivalent. See the references for the website to download and install R.

The code that follows can be copied and pasted into the R screen. Note that you will likely need to change the first line of the code to point to the folder location where you stored your data.

```
setwd ("c:/MyDocuments/EconEnvironAssess/Data") ## This depends on where
you saved your data; use the appropriate folder path
mydataset <- read.csv(file="GREENGOLD_Table.csv",head=TRUE,sep=",")
p1 <- princomp(~WRKDEATH + WRKHIINJ + WRKTOXIC + WRKHIDISEA + MXWKLYDISB
+ COVEMPINSU + STATPROTWR + UNEMP + UNEMPYTH + UNEMPDUR + EMP8593 +
OPPJOBWOME + OPPJOBMIN + AVEANNPAY + HHPOVERTY + GAPINCDIST + EDUCATTHS +
STTAXFAIR + BSTARTS + JOBGRTHNEW + HAZWASTE + TOTTOXCHEM + CANCERTOX +
SOLIDWASTE + SLDWSTRCYC + PESTCIDES + FERTILIZER + TOTWATER + SPILLS9092
+ GLOBWARM + AIRQUALITY + AVGMILEGAL + MILESDRIV + TOTBTU + CHGENRGYCO +
STSPNDIENV + STBUDGENV + ENVPOLICY + POLSUBINVE + EMISSJOBS, data =
mydataset, cor =TRUE)
p2 <- p1$scores
print(p1)
summary(p1) ## gives you the variance explained by each component
plot(p1) ## Scree plot
p3 <- cbind(p2, mydataset$STATE_FIPS2)
write.table(p3, file = "pcascores.tab") # open this in Excel and copy and
paste the state names as a column, delete the columns for the components
that are on the right tail of the scree plot, join this with the states
shapefile using STATE_FIPS2 to map component scores
loadings(p1)
```

Question 4: *Interpret the summary that appears below. How many components are needed to explain a significant portion of the variance?*



```
> summary(p1)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     3.0180653 2.3150008 2.0270597 1.67059103 1.62819990 1.32243848 1.27861889 1.21048076 1.11332576 1.01104053
Proportion of Variance 0.2277180 0.1339807 0.1027243 0.06977186 0.06627587 0.04372109 0.04087166 0.03663159 0.03098736 0.02555507
Cumulative Proportion  0.2277180 0.3616987 0.4644229 0.53419481 0.60047068 0.64419177 0.68506342 0.72169502 0.75268237 0.77823745
                          Comp.11    Comp.12    Comp.13    Comp.14    Comp.15    Comp.16    Comp.17    Comp.18    Comp.19
Standard deviation     1.00388385 0.92987562 0.88377314 0.87027898 0.81977827 0.77803397 0.72907576 0.72524045 0.66377395
Proportion of Variance 0.02519457 0.02161672 0.01952637 0.01893464 0.01680091 0.01513342 0.01328879 0.01314934 0.01101490
Cumulative Proportion  0.80343202 0.82504873 0.84457511 0.86350974 0.88031065 0.89544408 0.90873286 0.92188221 0.93289710
                          Comp.20     Comp.21     Comp.22     Comp.23    Comp.24     Comp.25     Comp.26     Comp.27     Comp.28
Standard deviation     0.65528402 0.615209269 0.540178401 0.518383686 0.46934550 0.443240538 0.418026247 0.374325502 0.354482623
Proportion of Variance 0.01073493 0.009462061 0.007294818 0.006718041 0.00550713 0.004911554 0.004368649 0.003502990 0.003141448
Cumulative Proportion  0.94363203 0.953094091 0.960388909 0.967106950 0.97261408 0.977525634 0.981894283 0.985397273 0.988538721
                           Comp.29     Comp.30     Comp.31     Comp.32      Comp.33      Comp.34      Comp.35      Comp.36
Standard deviation     0.322626416 0.294351767 0.247680238 0.243784436 0.1986919116 0.1765058951 0.1656314726 0.1385826994
Proportion of Variance 0.002602195 0.002166074 0.001533638 0.001485771 0.0009869619 0.0007788583 0.0006858446 0.0004801291
Cumulative Proportion  0.991140916 0.993306990 0.994840628 0.996326399 0.9973133607 0.9980922190 0.9987780636 0.9992581927
                            Comp.37      Comp.38      Comp.39      Comp.40
Standard deviation     0.1201505074 0.0837036964 0.0751212949 5.085891e-02
Proportion of Variance 0.0003609036 0.0001751577 0.0001410802 6.466571e-05
Cumulative Proportion  0.9996190963 0.9997942541 0.9999353343 1.000000e+00
```

Figure 1: Summary of components using R

Note that the first five components explain about 60 percent of the variance. Another indicator is the scree plot. Components after the point where the scree plot flattens out are usually ignored.

Question 5: *Where does the scree plot flatten out in the plot that follows?*

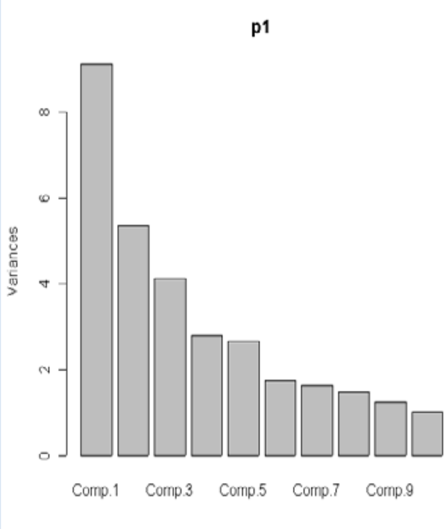
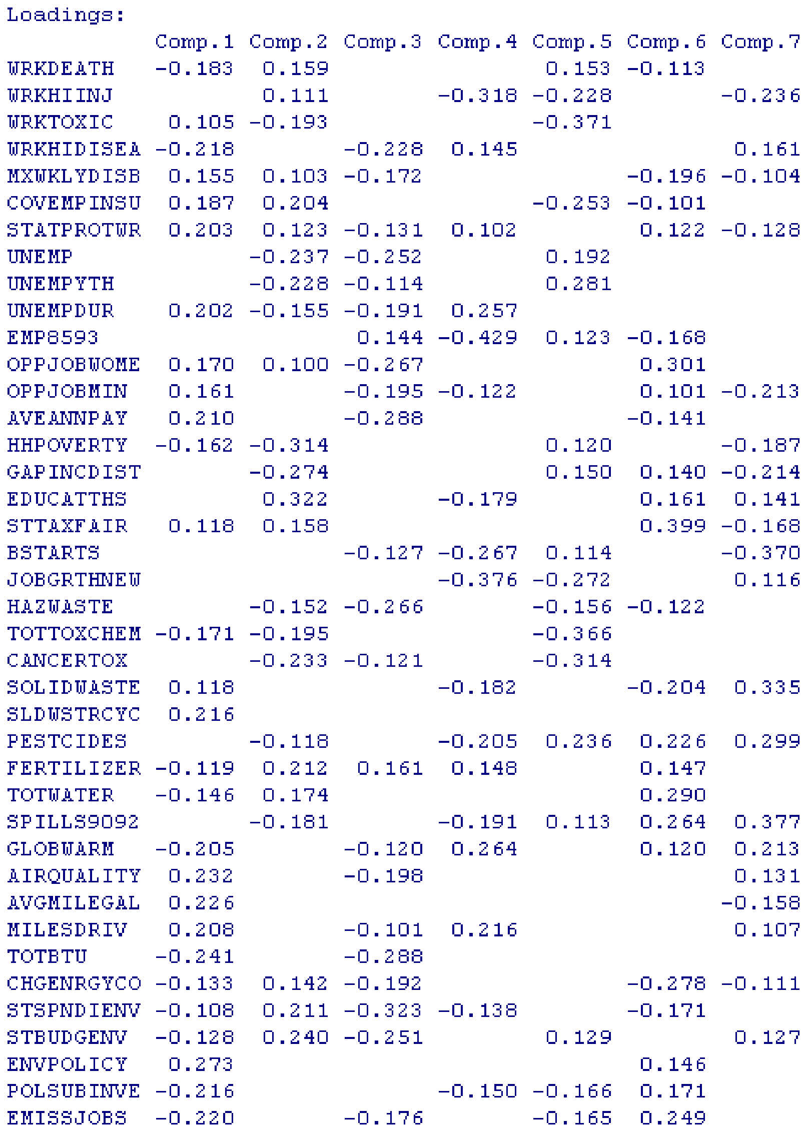


Figure 2: Scree plot of components using R

Question 6: *Interpret the loadings that you see in the R results (see the results in the table that follows). For example, for the first component, there is a (relatively) high correlation with average annual pay, solid-waste recycling, environmental policy, and low emissions per job—a possible indicator of both wealth and "environmental consciousness." Which states are likely to have high scores for this component?*

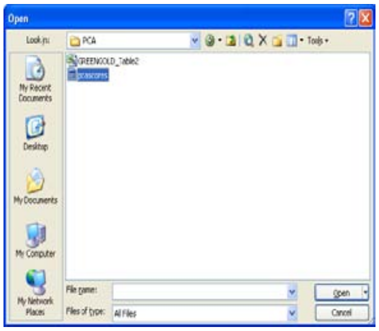
For example, component 4 loaded highest but with negative signs on the percentage of workers in high-injury industries (*WRKHIINJ*) and employment growth between 1985 and 1993 (*EMP8593*), which suggests that states with high scores on this component have high negative correlation with both variables. Likewise, states scoring high on component 2 have lower levels of poverty (*HHPOVERTY*) and higher numbers of educational attainment as measured by the percentage that graduated high school (*EDUCATTHS*).

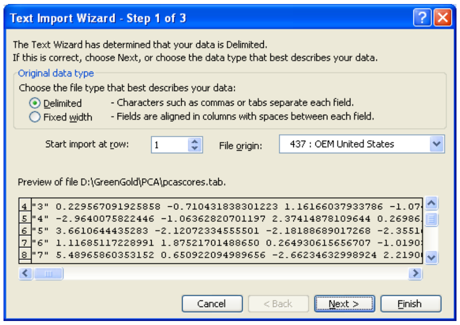


```
Loadings:
           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
WRKDEATH   -0.183  0.159                0.153 -0.113
WRKHIINJ           0.111        -0.318 -0.228        -0.236
WRKTOXIC    0.105 -0.193               -0.371
WRKHIDISEA -0.218        -0.228  0.145                0.161
MXWKLYDISB  0.155  0.103 -0.172               -0.196 -0.104
COVEMPINSU  0.187  0.204               -0.253 -0.101
STATPROTWR  0.203  0.123 -0.131  0.102         0.122 -0.128
UNEMP             -0.237 -0.252         0.192
UNEMPYTH          -0.228 -0.114         0.281
UNEMPDUR    0.202 -0.155 -0.191  0.257
EMP8593                   0.144 -0.429  0.123 -0.168
OPPJOBWOME  0.170  0.100 -0.267                0.301
OPPJOBMIN   0.161        -0.195 -0.122         0.101 -0.213
AVEANNPAY   0.210        -0.288               -0.141
HHPOVERTY  -0.162 -0.314                0.120        -0.187
GAPINCDIST        -0.274                0.150  0.140 -0.214
EDUCATTHS          0.322        -0.179         0.161  0.141
STTAXFAIR   0.118  0.158                       0.399 -0.168
BSTARTS                  -0.127 -0.267  0.114        -0.370
JOBGRTHNEW                      -0.376 -0.272         0.116
HAZWASTE          -0.152 -0.266        -0.156 -0.122
TOTTOXCHEM -0.171 -0.195               -0.366
CANCERTOX         -0.233 -0.121        -0.314
SOLIDWASTE  0.118               -0.182        -0.204  0.335
SLDWSTRCYC  0.216
PESTCIDES         -0.118        -0.205  0.236  0.226  0.299
FERTILIZER -0.119  0.212  0.161  0.148         0.147
TOTWATER   -0.146  0.174                       0.290
SPILLS9092        -0.181        -0.191  0.113  0.264  0.377
GLOBWARM   -0.205        -0.120  0.264         0.120  0.213
AIRQUALITY  0.232        -0.198                       0.131
AVGMILEGAL  0.226                                    -0.158
MILESDRIV   0.208        -0.101  0.216                0.107
TOTBTU     -0.241        -0.288
CHGENRGYCO -0.133  0.142 -0.192               -0.278 -0.111
STSPNDIENV -0.108  0.211 -0.323 -0.138        -0.171
STBUDGENV  -0.128  0.240 -0.251         0.129         0.127
ENVPOLICY   0.273                              0.146
POLSUBINVE -0.216               -0.150 -0.166  0.171
EMISSJOBS  -0.220        -0.176        -0.165  0.249
```

Figure 3: Loadings of variables on components

3 Before you can join this data to the shapefile, you need to clean it up. Open the results of the PCA *pcascores.tab* in Excel. See the screen captures that follow to see how you could read this as a tab- and space-delimited table:




Text Import Wizard - Step 1 of 3

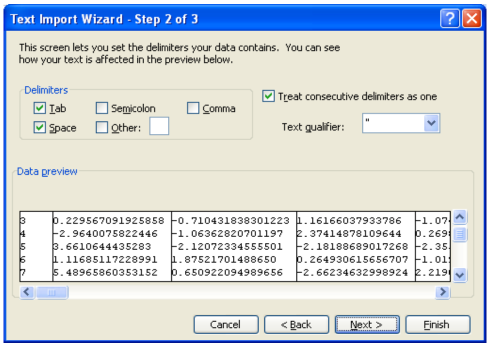
The Text Wizard has determined that your data is Delimited.
If this is correct, choose Next, or choose the data type that best describes your data.

Original data type
Choose the file type that best describes your data:
● Delimited - Characters such as commas or tabs separate each field.
○ Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row: 1 File origin: 437 : OEM United States

Preview of file D:\GreenGold\PCA\pcascores.tab.

```
4 "3" 0.229567091925858 -0.710431838301223 1.1616603793378 6 -1.07
5 "4" -2.9640075822446 -1.06362820701197 2.37414878109644 0.26986
6 "5" 3.6610644435283 -2.12072334555501 -2.18188689017268 -2.3551
7 "6" 1.11685117228991 1.87521701488650 0.264930615656707 -1.0190
8 "7" 5.48965860353152 0.650922094989656 -2.66234632998924 2.2190
```


Text Import Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

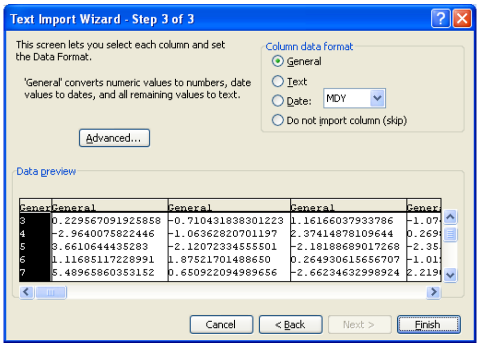
Delimiters
☑ Tab ☐ Semicolon ☐ Comma
☑ Space ☐ Other:

☑ Treat consecutive delimiters as one

Text qualifier: "

Data preview

```
3  0.229567091925858  -0.710431838301223  1.16166037933786   -1.07
4  -2.9640075822446   -1.06362820701197   2.37414878109644   0.269
5  3.6610644435283    -2.12072334555501   -2.18188689017268  -2.35
6  1.11685117228991   1.87521701488650    0.264930615656707  -1.01
7  5.48965860353152   0.650922094989656   -2.66234632998924  2.219
```


Text Import Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

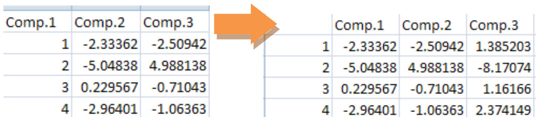
'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Column data format
● General
○ Text
○ Date: MDY
○ Do not import column (skip)

Data preview

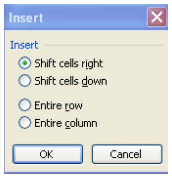
```
Gener General            General             General            Gener
3     0.229567091925858  -0.710431838301223  1.16166037933786   -1.07
4     -2.9640075822446   -1.06362820701197   2.37414878109644   0.269
5     3.6610644435283    -2.12072334555501   -2.18188689017268  -2.35
6     1.11685117228991   1.87521701488650    0.264930615656707  -1.01
7     5.48965860353152   0.650922094989656   -2.66234632998924  2.219
```

4 The first column will be incorrectly labeled as *Comp. 1*. The first column is just an ID indicating the row (state). Shift the titles by one cell to the right using *Insert Cells » Shift cells right* and type a new name for the first column, such as *ID*. The last column, which has no name, should be called *ST_FIPS2*.



Comp.1	Comp.2	Comp.3
1	-2.33362	-2.50942
2	-5.04838	4.988138
3	0.229567	-0.71043
4	-2.96401	-1.06363

	Comp.1	Comp.2	Comp.3
1	-2.33362	-2.50942	1.385203
2	-5.04838	4.988138	-8.17074
3	0.229567	-0.71043	1.16166
4	-2.96401	-1.06363	2.374149


Insert

Insert
● Shift cells right
○ Shift cells down
○ Entire row
○ Entire column

5 Save it as an Excel file after deleting the columns for the components after Comp. 10 (components after this explain marginally smaller amounts of the variance). However, keep the column named *ST_FIPS2*.

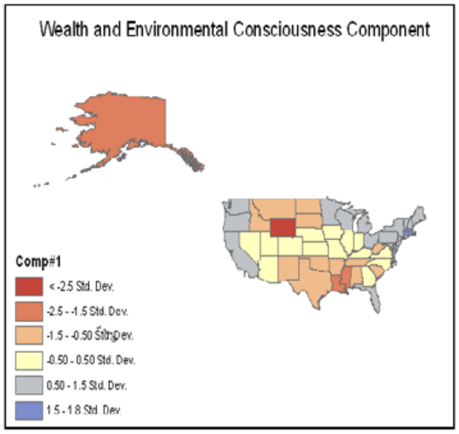
6 Add this new Excel file in ArcMap. Join it to the shapefile using *ST_FIPS2* as the key attribute.

VISUALIZE

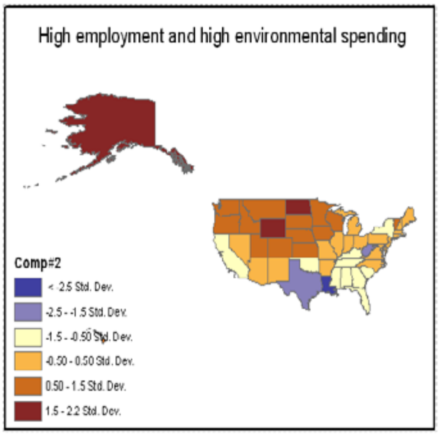
Maps 2a and 2b show the spatial patterns of the first two components. What do these patterns indicate about the components? Note that the titles shown for the components are interpretations of the PCA output. While there are general guides for the interpretation, the actual application depends on the output. In general, there is no assurance that components can be described by only two of the original variables. Moreover, none of the terms *wealth*, *environmental consciousness*, *high employment*, or *high environmental spending* actually appears in the PCA output—only the abbreviated variable names appear.

Also note that the classification scheme here is the standard deviational units. The PCA results are output as normalized scores, and so the use of such a classification scheme is appropriate. Scores may also be classified using other schemes. Here, the interpretation is that darker colors on both ends (both negative and positive standard deviational units) of the scheme are different from the average values.

You could use the maps below to compare regions. For example, it is possible that the states scoring high on component 1 and component 2 (New England states) have perhaps exported their environmentally hazardous industries to those that score low on component 1 and high on component 2 (mountain states). Also, notice that there is much more variation in component 1 in the southeastern states than there is in component 2. What does this say about environmental and economic development policy in the southeastern states?



Map 2a: Spatial variation of component 1



Map 2b: Spatial variation of component 2

Question 7: *Create maps of components that are interesting. (In the table of contents, right-click the layer and then click* Properties » Symbology » Quantities *and the appropriate components for* Values*).What kinds of patterns do you observe?*

Question 8: *Test if there are regional or global clusters in these components using Moran's I and Anselin* statistics *in the* Spatial Statistics *toolbox. Show where the hot spots are (if any) for various components (extra credit).*

Submit your work

Submit answers to questions 1–8 along with relevant screen captures. Describe how these maps and statistics could inform environmental and economic policy.

Credits

Data

Data used in this activity courtesy of ArcUSA, US Census, and Esri, *Esri Data & Maps*, 2008.

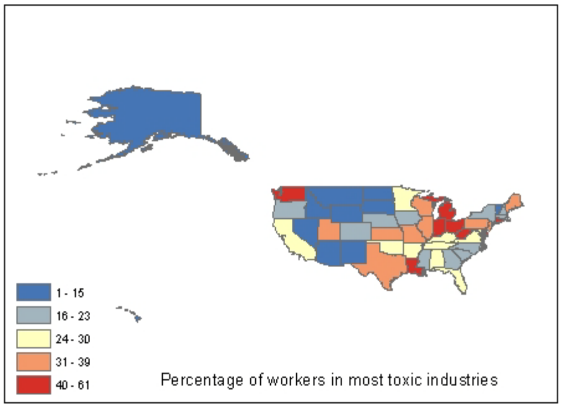
Instructor resources

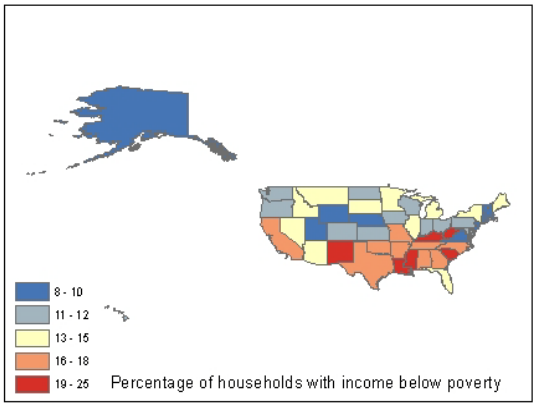
Additional information

This exercise is intended for the student who wishes to combine statistical and spatial methods. The steps shown here can be extended to the use of the results of other statistical methods in R within ArcGIS. This lab was used in an environmental assessment class to understand whether green indicators correlate with wealth in the United States at the state level. Modifiable areal unit problem (MAUP) issues need to be addressed, and intrastate differences need to be discussed as well. Principal component analysis is used in remote sensing as well, so the instructor could combine this with a lecture on classification of satellite data. The question on Moran's I and local Moran can be omitted if the students have not studied spatial statistics.

Answers to questions

Question 1: *Do high percentages of workers in toxic industries and high poverty levels occur in the same states? Where are these variables spatially clustered?*





Answer: The patterns above suggest that the highest percentage of workers in the most toxic industries are in the Midwestern states of Michigan, Ohio, Indiana, and West Virginia, as well as in Louisiana and Washington. Overall, the higher values cluster in the Midwest down to the South in Texas and Louisiana. This pattern of clustering does not coincide exactly with the states with the highest poverty, but several states in the South, like Louisiana and Texas, as well as some in the Midwest, do have among the higher percentages of households living in poverty.

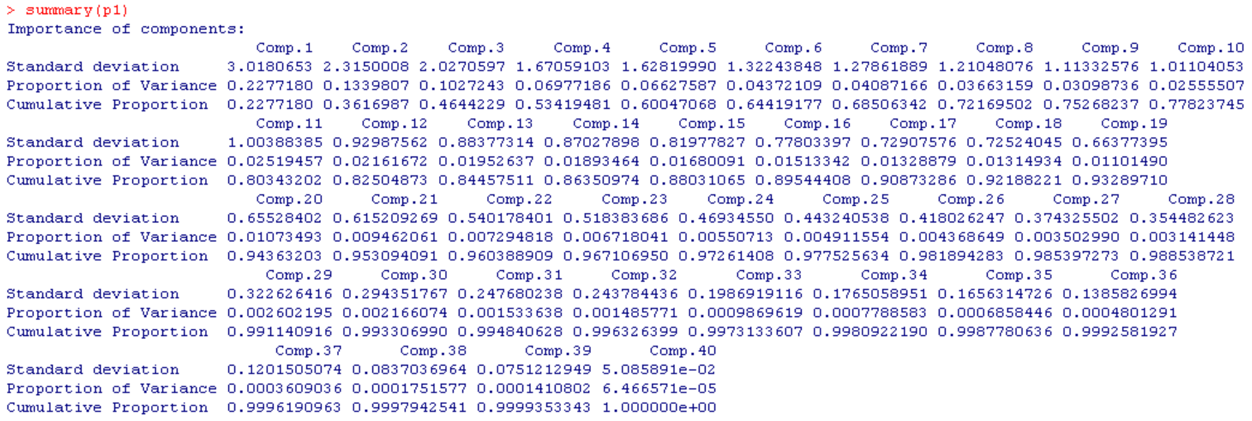
Question 2: *Choose some other economic development variables and environmental variables, and then conjecture as to whether they are likely to be clustered spatially. Use maps to confirm or rebut your conjectures. How easy or difficult is it to confirm your hypotheses?*

Answer: In this answer, students should show evidence of having examined the Microsoft Excel spreadsheet for environmental variables such as toxic spills in 1990–92 (SPILLS9092) versus economic indicators such as average annual pay (AVEANNPAY). It could be speculated, for example, that southern states will have higher numbers of toxic spills but lower average annual pay. Their maps will find that this is not the case and that, in fact, spills appear to often be in high-income states such as California.

Question 3: *Report the Moran's I for several variables of interest.*

Answer: This answer should be based on what students find in questions 1 and 2. They should be encouraged to look for patterns of clustering. For example, AVEANNPAY has a Moran's I of 0.17, and the report shows that this indicates significant clustering. AIRQUALITY has a Moran's I of 0.22 and significant clustering. However, SOLIDWASTE and BSTARTS (business starts) show patterns that are not significantly different from random.

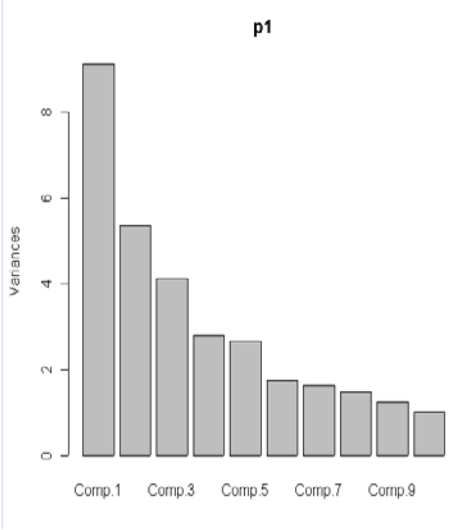
Question 4: *Interpret the summary that appears below. How many components are needed to explain a significant portion of the variance?*



```
> summary(p1)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     3.0180653 2.3150008 2.0270597 1.67059103 1.62819990 1.32243848 1.27861889 1.21048076 1.11332576 1.01104053
Proportion of Variance 0.2277180 0.1339807 0.1027243 0.06977186 0.06627587 0.04372109 0.04087166 0.03663159 0.03098736 0.02555507
Cumulative Proportion  0.2277180 0.3616987 0.4644229 0.53419481 0.60047068 0.64419177 0.68506342 0.72169502 0.75268237 0.77823745
                          Comp.11    Comp.12    Comp.13    Comp.14    Comp.15    Comp.16    Comp.17    Comp.18    Comp.19
Standard deviation     1.00388385 0.92987562 0.88377314 0.87027898 0.81977827 0.77803397 0.72907576 0.72524045 0.66377395
Proportion of Variance 0.02519457 0.02161672 0.01952637 0.01893464 0.01680091 0.01513342 0.01328879 0.01314934 0.01101490
Cumulative Proportion  0.80343202 0.82504873 0.84457511 0.86350974 0.88031065 0.89544408 0.90873286 0.92188221 0.93289710
                          Comp.20     Comp.21     Comp.22     Comp.23    Comp.24     Comp.25     Comp.26     Comp.27     Comp.28
Standard deviation     0.65528402 0.615209269 0.540178401 0.518383686 0.46934550 0.443240538 0.418026247 0.374325502 0.354482623
Proportion of Variance 0.01073493 0.009462061 0.007294818 0.006718041 0.00550713 0.004911554 0.004368649 0.003502990 0.003141448
Cumulative Proportion  0.94363203 0.953094091 0.960388909 0.967106950 0.97261408 0.977525634 0.981894283 0.985397273 0.988538721
                           Comp.29     Comp.30     Comp.31     Comp.32      Comp.33      Comp.34      Comp.35      Comp.36
Standard deviation     0.322626416 0.294351767 0.247680238 0.243784436 0.1986919116 0.1765058951 0.1656314726 0.1385826994
Proportion of Variance 0.002602195 0.002166074 0.001533638 0.001485771 0.0009869619 0.0007788583 0.0006858446 0.0004801291
Cumulative Proportion  0.991140916 0.993306990 0.994840628 0.996326399 0.9973133607 0.9980922190 0.9987780636 0.9992581927
                            Comp.37      Comp.38      Comp.39      Comp.40
Standard deviation     0.1201505074 0.0837036964 0.0751212949 5.085891e-02
Proportion of Variance 0.0003609036 0.0001751577 0.0001410802 6.466571e-05
Cumulative Proportion  0.9996190963 0.9997942541 0.9999353343 1.000000e+00
```

Answer: Note that the first five components explain about 60 percent of the variance (add the proportion of variance) or look at the cumulative proportion. The fifth component explains about 6.6 percent compared to the first, which explains about 22.8 percent. The components that follow explain smaller and smaller portions of the variance in the data, so you can ignore them.

Question 5: *Where does the scree plot flatten out in the plot that follows?*



Answer: The scree plot begins to flatten out after component 1, but you could argue that the first three components may still be useful (there is an even steeper fall after component 3).

Question 6: *Interpret the loadings that you see in the R results (see the results in the table that follows). For example, for the first component, there is a (relatively) high correlation with average annual pay, solid-waste recycling, environmental policy, and low emissions per job—a possible indicator of both wealth and "environmental consciousness." Which states are likely to have high scores for this component?*

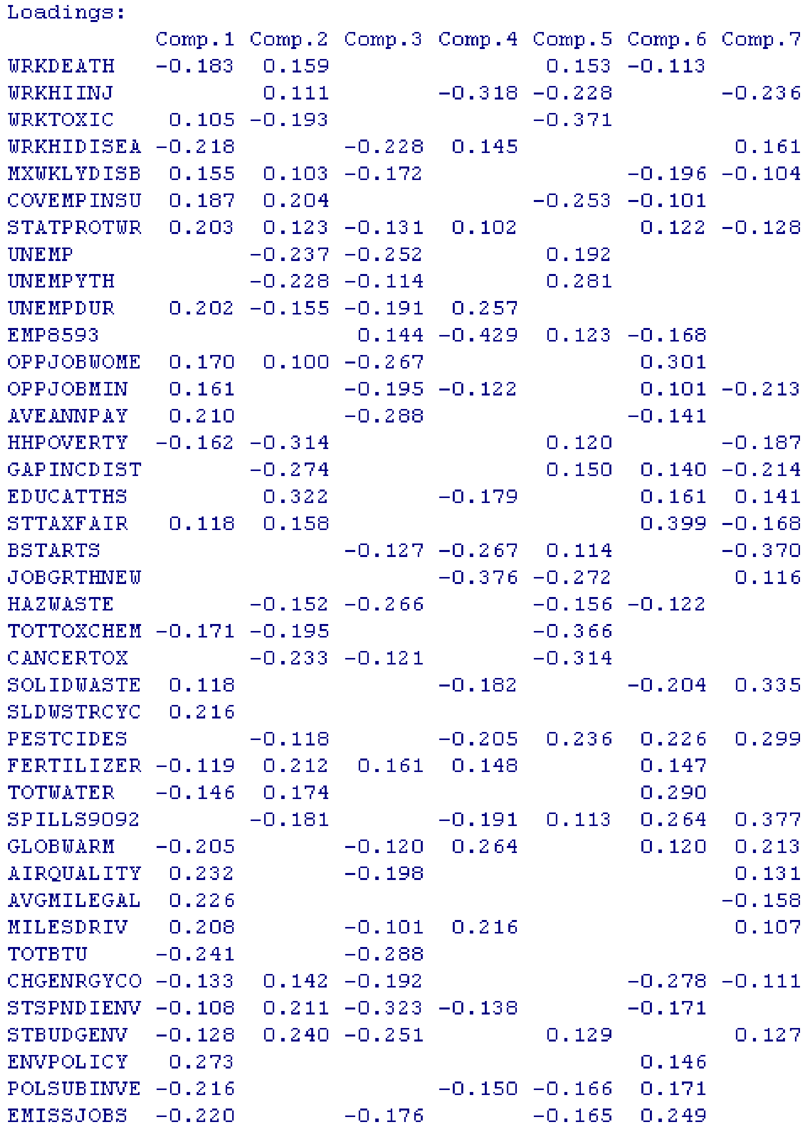
Answer: For example, component 4 loaded highest but with negative signs on the percentage of workers in high-injury industries (WRKHIINJ) and employment growth between 1985 and 1993 (EMP8593), which suggests that states with high scores on this component have high negative correlation with both variables.

States scoring high on component 2 have lower levels of poverty (HHPOVERTY) and higher levels of educational attainment as measured by the percentage that graduated high school (EDUCATTHS) and low unemployment (UNEMP) as well as high spending on the environment (STBUDGENV). This may indicate the coastal states, but it could also indicate mountain states that have low unemployment.

Component 1 included the following loadings that were relatively high (over 0.2): average annual pay, unemployment duration, statutory protection for workers, and an environmental protection record.

The component was also negatively correlated with total BTU (energy consumption), emissions per job, and workers in industries at high risk for disease. This suggests a component for which high scores might indicate wealth as well as environmental consciousness. Perhaps this would mean the coastal states in the Northeast and on the West Coast would stand out in component 1.

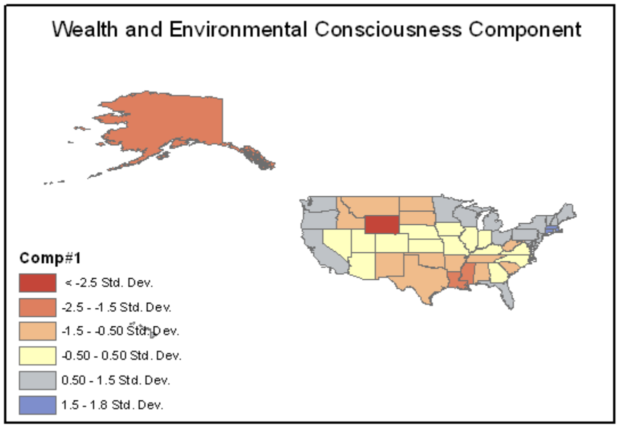
Component 3 is a mixed bag in that it is negatively correlated with average annual pay, low unemployment, opportunities for women, average annual pay (economic "bad"), and state budget environmental spending (environmental "bad") but also negatively correlated with hazardous-waste releases and workers in disease-risk industries (environmental "good").

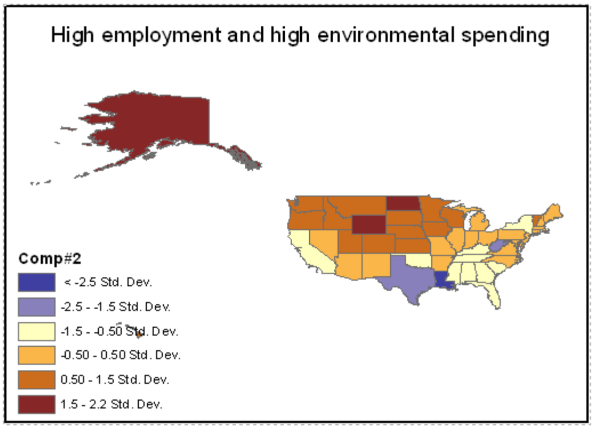


```
Loadings:
           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
WRKDEATH   -0.183  0.159                0.153 -0.113
WRKHIINJ           0.111        -0.318 -0.228        -0.236
WRKTOXIC    0.105 -0.193               -0.371
WRKHIDISEA -0.218        -0.228  0.145                0.161
MXWKLYDISB  0.155  0.103 -0.172               -0.196 -0.104
COVEMPINSU  0.187  0.204               -0.253 -0.101
STATPROTWR  0.203  0.123 -0.131  0.102         0.122 -0.128
UNEMP             -0.237 -0.252         0.192
UNEMPYTH          -0.228 -0.114         0.281
UNEMPDUR    0.202 -0.155 -0.191  0.257
EMP8593                   0.144 -0.429  0.123 -0.168
OPPJOBWOME  0.170  0.100 -0.267                0.301
OPPJOBMIN   0.161        -0.195 -0.122         0.101 -0.213
AVEANNPAY   0.210        -0.288               -0.141
HHPOVERTY  -0.162 -0.314                0.120        -0.187
GAPINCDIST        -0.274                0.150  0.140 -0.214
EDUCATTHS          0.322        -0.179         0.161  0.141
STTAXFAIR   0.118  0.158                       0.399 -0.168
BSTARTS                  -0.127 -0.267  0.114        -0.370
JOBGRTHNEW                      -0.376 -0.272         0.116
HAZWASTE          -0.152 -0.266        -0.156 -0.122
TOTTOXCHEM -0.171 -0.195               -0.366
CANCERTOX         -0.233 -0.121        -0.314
SOLIDWASTE  0.118               -0.182        -0.204  0.335
SLDWSTRCYC  0.216
PESTCIDES         -0.118        -0.205  0.236  0.226  0.299
FERTILIZER -0.119  0.212  0.161  0.148         0.147
TOTWATER   -0.146  0.174                       0.290
SPILLS9092        -0.181        -0.191  0.113  0.264  0.377
GLOBWARM   -0.205        -0.120  0.264         0.120  0.213
AIRQUALITY  0.232        -0.198                       0.131
AVGMILEGAL  0.226                                    -0.158
MILESDRIV   0.208        -0.101  0.216                0.107
TOTBTU     -0.241        -0.288
CHGENRGYCO -0.133  0.142 -0.192               -0.278 -0.111
STSPNDIENV -0.108  0.211 -0.323 -0.138        -0.171
STBUDGENV  -0.128  0.240 -0.251         0.129         0.127
ENVPOLICY   0.273                              0.146
POLSUBINVE -0.216               -0.150 -0.166  0.171
EMISSJOBS  -0.220        -0.176        -0.165  0.249
```

Question 7: *Create maps of the components that are interesting. (In the table of contents, right-click the layer and then click* Properties » Symbology » Quantities *and the appropriate components for* Values*).What kinds of patterns do you observe?*

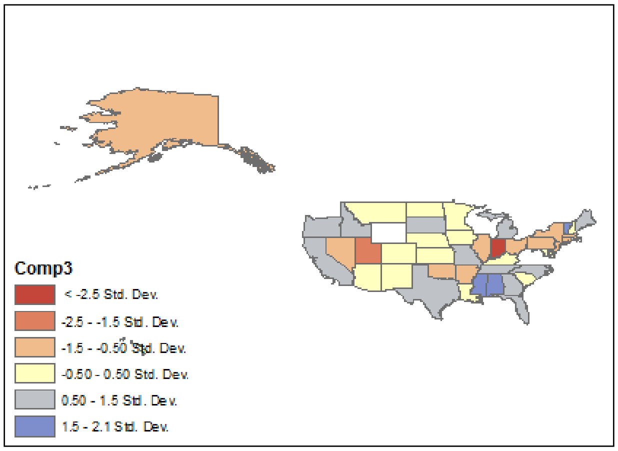
Answer: Students should be expected to map, at the most, four components.



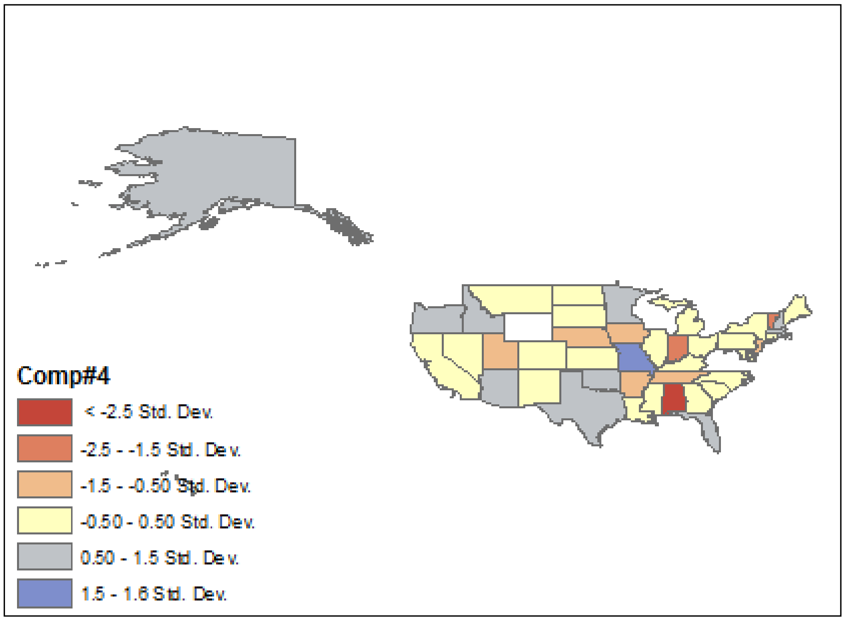


Component 1 indicates high scores in the Northeast, Florida, and West Coast states, as one would expect, indicating wealth and environmental consciousness.

Component 2 shows high scores in the mountain states, Alaska, and the Northeast, so this component shows a very different indicator than component 1. These are perhaps states that have high employment and environmental spending for very different reasons (the mountain states stand out as the highest values, and the southern states have the lowest values).



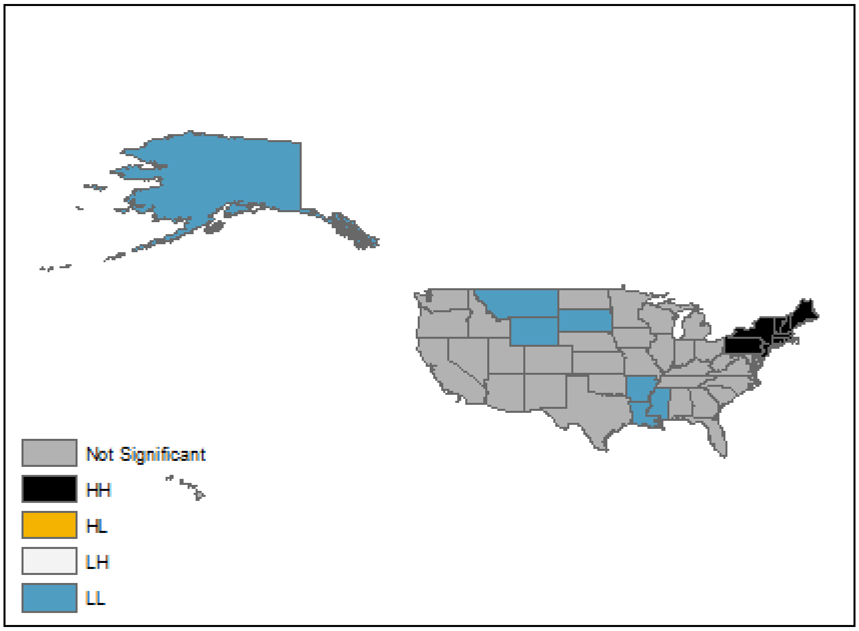
This component pulls together very different states that include the wealthy, such as California, and least wealthy, such as Louisiana, giving them similar scores in a component that was found earlier to be a mixed bag.



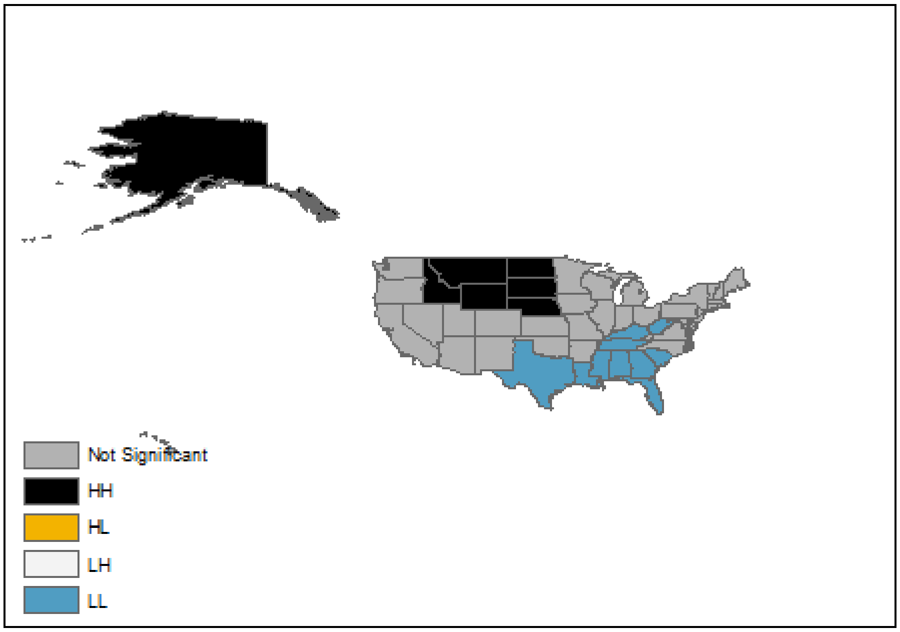
It is interesting to see that states with high-risk jobs show up in very different parts of the country. It does appear to indicate that wealthier states have fewer such jobs but also lower job growth, perhaps.

Question 8: *Test if there are regional or global clusters in these components using Moran's I and Anselin statistics in the* Spatial Statistics *toolbox. Show where the hot spots are (if any) for various components (extra credit).*

Answer: The first component had the highest Moran's I of 0.27, while the second and fourth components are somewhat less clustered but still significantly so (about 0.13 Moran's I for both). The third component—the mixed bag—was a random pattern.



For component 1, there is a significant cluster of high values surrounded by other high values in the Northeast. There is a cluster of low values in the South and in the North. Here, HH indicates high surrounded by high, LL indicates low surrounded by low, HL is high surrounded by low values, and LH is the opposite.



For component 2, the high values are in the mountain states, and the low values are in the southern states.

References and further reading

Fotheringham, A. Stewart, Chris Brunsdon, and Martin Charlton. 2000*. Quantitative Geography: Perspectives on spatial data analysis*. London: Sage Publications.

R documentation

See `http://www.r-project.org/` to download and install the R software (binaries are available for Windows, Linux, and Mac).

Documentation of princomp
`http://stat.ethz.ch/R-manual/R-patched/library/stats/html/princomp.html`

Be aware that there are several different functions in R for PCA, which have slight but important differences. Likewise, if the PCA option is used in some other software package, there may appear to be differences in the output, even though the underlying theory is the same.

A search on Google for "R tutorial" and "R tutorial video" will yield many useful and helpful links.

Web resources on PCA

http://www.statsoft.com/textbook/principal-components-factor-analysis/

http://www.fon.hum.uva.nl/praat/manual/Principal_component_analysis.html

http://cran.r-project.org/web/packages/HSAUR/vignettes/Ch_principal_components_analysis.pdf

Spatial statistics resources for the testing of clustering

Detecting hot spots using cluster analysis and GIShttp://www.tonygrubesic.net/hot_spot.pdf

Spatial cluster analysis
http://www.terraseer.com/pdf/jacquez_ch22_preprint.pdf

Data source

The data file was compiled by students at Harvard University. A standard shapefile of the state boundaries from Esri was used.