Educational Performance and Family Income

Diamonds on the soles of scholarship?

— by Jeff Blossom

Introduction

Problem

Two young girls entering elementary school in Boston aspire to be doctors. Both come from two-parent, one-income families with two siblings. In 13 years and with hard work, assuming all other variables equal, how will their SAT scores compare given that one comes from an impoverished household and the other from a household that is well-to-do?

Is there a relationship between educational performance and family income in the State of Massachusetts, and how might that relationship be illustrated?

Using geographic information systems (GIS), it is possible to combine academic test score information by school district with family income data gathered from the US Census to make a series of maps that illustrate income and educational performance across the state, and then analyze the geographic patterns of these maps.

Location

The Commonwealth of Massachusetts, USA

Time to complete the lab

Two hours

Keywords: choropleth, statistics, data classification, correlation, spatial join, overlay analysis

Prerequisites

Cursory familiarity with ArcGIS software

Data used in this lab

- School district boundaries, town boundaries, test scores, census data
- Geographic coordinate system: NAD 1983
- Datum: NAD 1983
- Projected coordinate system: Massachusetts Mainland State Plane (meters)

Student activity

The correlation between quality of education and family income in Massachusetts cities and towns is well known, and this topic is often discussed among policy makers and citizens. The Massachusetts Comprehensive Assessment System (MCAS) provides test score information by school district, and the Office of Geographic Information (MassGIS) provides town boundaries, school district boundaries, and other geographic data. You will combine this data with census family income data to make a series of maps that illustrate income and educational performance across the state. Given these maps, you will be able to analyze geographic patterns and examine the utility of using maps to display these phenomena.

A *choropleth* map is one that shades different geographic areas by a value or statistic. Demographic data such as population density, average age, racial distribution, and income per area is often portrayed using choropleth maps. It is a form of communication that can be effective at revealing areas of similar, different, or outlying characteristics. In this exercise, you'll make three choropleth maps of Massachusetts (MA) school districts. One will illustrate the landscape of family income in Massachusetts. Another will illustrate educational performance in Massachusetts. The third map will illustrate four correlations: high income/high educational performance (positive correlation), low income/low educational performance (positive correlation), high income/low educational performance (negative correlation), and low income/high educational performance (negative correlation).

In this exercise, you will do the following:

Prepare the data

- Copy the school districts, town boundaries, and census data into your project folder.

Create maps

- Median family income
- MCAS test score achievement (10th grade for 2010)
- Educational achievement/income correlation maps

Analyze results

- Analyze the spatial relationship between towns and school districts in Massachusetts.
- Explore the relationship between family income and school test scores.

PREPARE YOUR WORKSPACE

1 Create a *SpatiaLABS* folder under the C:\ folder and an *MA_Education* subfolder.

2 Locate MA_Education.zip and extract it into the C:\SpatiaLABS\MA_Education workspace.

3 Examine all data in the MA_Education folder. You should have Town_Boundaries, School_Districts, and Block_Groups shapefiles.

COMPARE TOWNS, SCHOOL DISTRICTS, AND BLOCK GROUPS IN MASSACHUSETTS

1 Start ArcMap and add the *Town_Boundaries* shapefile. This shapefile contains locations of all 351 towns and cities in Massachusetts. Add the *School_Districts* shapefile. This contains boundaries of high school districts in Massachusetts, downloaded from MassGIS (the state GIS clearinghouse).

Question 1: *How many high school districts are there in Massachusetts?*

Question 2: *Describe the spatial relationship between towns and school districts. Which one contains larger areas on average? Are the boundary lines between towns and school districts coincident?*

2 Add the *Block_Groups* dataset.

Question 3: *What are census block groups? Is there a small variation or large variation in the size of a census block group? Use research from the web and cite your information sources.*

Attributes of the *Block_Groups* dataset are as follows:

ID: the unique identifier for each block group

POP: total population from 2010 in the block group

MHI: median household income from 2011 in the block group

3 Examine block group population in Massachusetts using visual overlay.

4 Open the *Block_Groups* attribute table.

5 Right-click the *POP* field, and then click *Sort » Descending*.

6 Select the block group with the highest population by clicking the gray tab on the left of the row. Click the *Zoom to Selected* button. Examine where this block group is compared to the *Town_Boundaries* layer and the *Imagery Basemap* to help answer the following:

Question 4: *The most populous block group is in what town/city?*

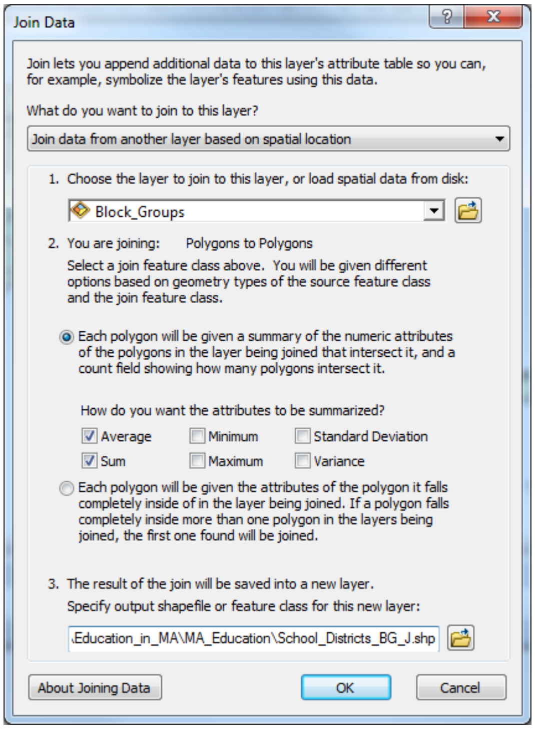
Question 5: *Name two of the towns/cities that have a block group with zero population. Using the Imagery Basemap, describe what you see in these zero-population block groups.*

SUMMARIZE INCOME BY SCHOOL DISTRICTS AND MAKE A CHOROPLETH MAP

To start analyzing correlations between family income and MCAS test scores, having a choropleth map that shows median family income by school district will be helpful. At this point, the school districts layer contains only the school district name and MCAS test performance. To make a map of median family income by school district, it will be necessary to aggregate the block group median family income values by school district. The command used to perform this type of overlay analysis between two different map layers is called "spatial join." It is a very useful command when aggregating information from two different data layers.

1 Perform the spatial join by right-clicking the *School_Districts* map layer and clicking *Joins and Relates » Join*.

2 In the spatial join window, specify *Join data from another layer based on spatial location* as the join type. Choose the *Block_Groups* layer to join to. Under *How do you want the attributes to be summarized?*, click the check boxes next to *Average* and *Sum*. Specify the output file as ***School_Districts_BG_J.*** Your screen should look like the figure below.



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 data from another layer based on spatial location

1. Choose the layer to join to this layer, or load spatial data from disk:

Block_Groups

2. You are joining: Polygons to Polygons

Select a join feature class above. You will be given different options based on geometry types of the source feature class and the join feature class.

● Each polygon will be given a summary of the numeric attributes of the polygons in the layer being joined that intersect it, and a count field showing how many polygons intersect it.

How do you want the attributes to be summarized?

☑ Average ☐ Minimum ☐ Standard Deviation

☑ Sum ☐ Maximum ☐ Variance

○ Each polygon will be given the attributes of the polygon it falls completely inside of in the layer being joined. If a polygon falls completely inside more than one polygon in the layers being joined, the first one found will be joined.

3. The result of the join will be saved into a new layer.

Specify output shapefile or feature class for this new layer:

\Education_in_MA\MA_Education\School_Districts_BG_J.shp

3 Click *OK* to run the spatial join.

When complete, *School_Districts_BG_J* will be added to the map.

4 Open the *School_Districts_BG_J* attribute table.

5 Note the *Avg_MHI* attribute. This contains the average median household income for all the block groups in each school district. Right-click this field and click *Statistics*.

Question 6: *What are the mean, minimum, and maximum median household incomes for school districts in Massachusetts? What is the variation between the min. and max. values? Is this disparity (or lack of) surprising?*

Now it is time to make the choropleth map of median household income by school district.

In mapping a statistic, choropleth maps can appear drastically different based on the choices of the cartographer. When making a choropleth map, you should be thinking about these cartographic principles:

- Data classification technique
- Number of data classes
- Colors used

Next, you will learn what these principles are and experiment with different ways to use them.

6 Right-click the *School_Districts_BG_J* layer, click *Properties,* and then click the *Symbology* tab.

7 Click the *Quantities* option on the left of the *Symbology* dialog box and select the *Avg_MHI* field as the *Value Field* to symbolize the map by.

When symbolizing choropleth maps, it is important to consider the data classification type. This determines which bins, or groups, the data is organized into for display on the map.

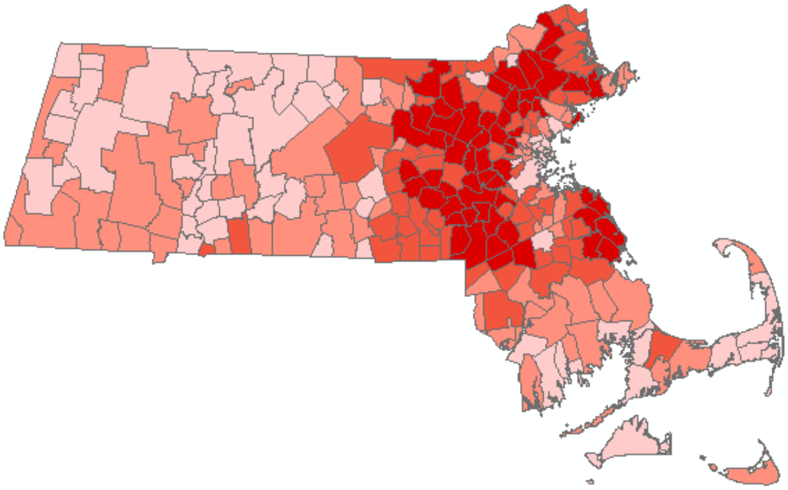
8 Note that the classification type defaults to Natural Breaks (Jenks) with 5 classes. To change this, click the *Classify* button. Examine the distribution of median income values in the histogram. Try all the different classification methods to see how each one changes the map.

Question 7: *What classification technique might be the best for identifying outliers in the dataset (i.e., either extremely high or extremely low data values)?*

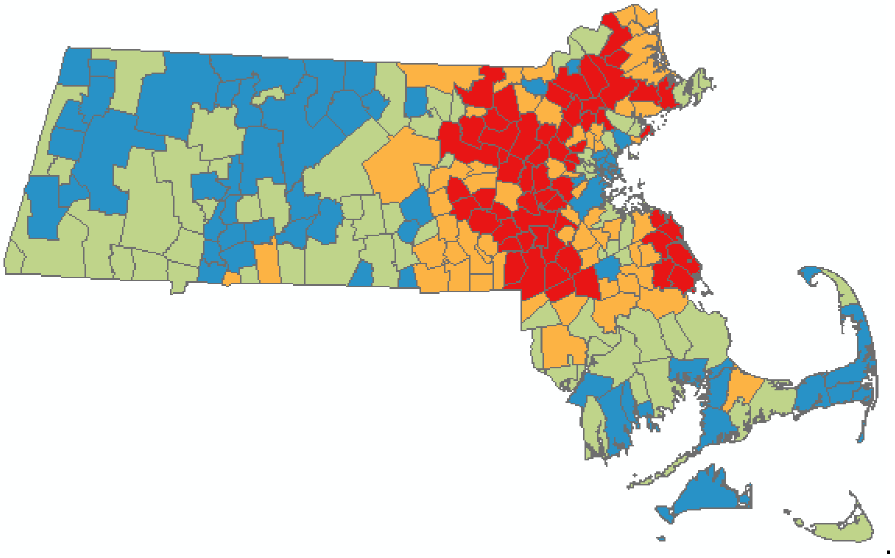
9 The number of data classes your data is grouped into also affects how the resulting map looks. Experiment with changing the number of data classes and how it changes the look of the resulting map.

Question 8: *What effect does increasing the number of classes have on the map? What effect does decreasing the number of classes have on the map?*

One goal of the cartographer is to use symbology and colors on the map that reflect the type of phenomena being portrayed. Certain color schemes can affect how the map is understood. One principle to try and follow when using ratio data is to use varying values (lightness) of the same hue (color) to depict low to high values. The graphic of school districts in Massachusetts shaded by median family income depicts this using the color red, as shown in the figure below.



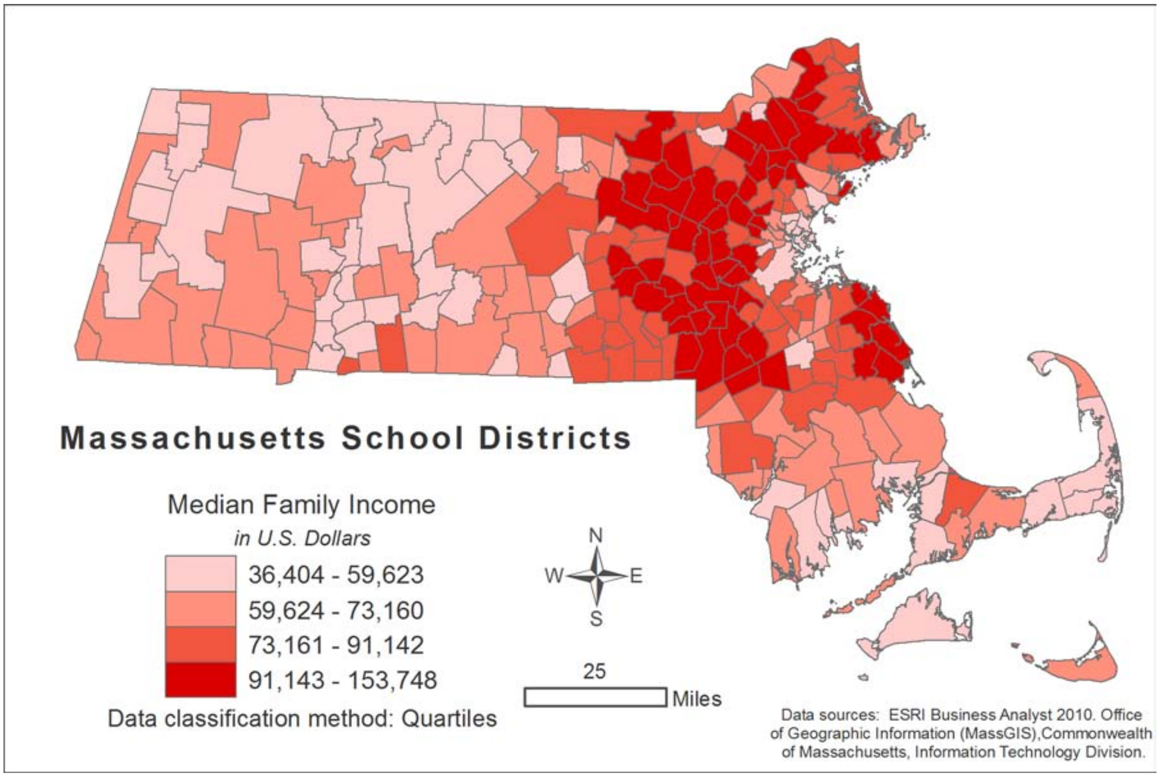
Using the same color suggests a natural change from low to high. Displaying this data with the same classification scheme but different colors is presented in the figure below.



On this map, the red stands out, but so does the blue. So it may be useful in focusing the map reader's attention on the high- and low-income areas, if that is the purpose of the map. Blue is typically associated with cold, or low, and red is often associated with hot, or high. The green and yellow are difficult to naturally interpret, as neither color is normally associated with high or low, or related to the red and blue.

Question 9: *What is a "diverging color scheme"? With what kind of data would this scheme be used?*

10 Now make your school districts median household income choropleth map. Make sure to include a title, legend, scale bar, and data sources reference (the block group MHI data is from the Esri Business Analyst 2010 dataset). Also note the data classification technique used on the map. Doing this is a way of giving "full disclosure" to the map viewer in terms of communicating exactly how the data was classified. Here's an example in the figure below.


Massachusetts School Districts
Median Family Income
in U.S. Dollars
36,404 - 59,623
59,624 - 73,160
73,161 - 91,142
91,143 - 153,748
Data classification method: Quartiles
N
W
E
S
25
Miles
Data sources: ESRI Business Analyst 2010. Office of Geographic Information (MassGIS),Commonwealth of Massachusetts, Information Technology Division.

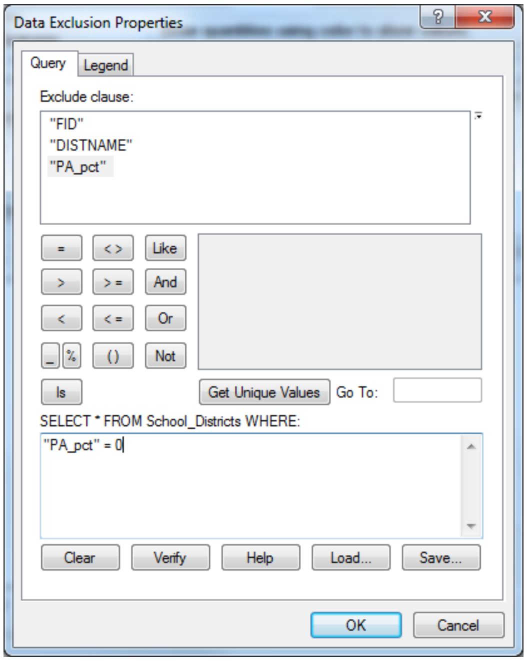
Map 1. Median family income by school district.

Question 10: *Examine the map. What is the spatial pattern of family income in Massachusetts?*

MAP TEST SCORE ACHIEVEMENT

MCAS tests grades 3-8 and grade 10 in math and English. The results from these tests are reported on an individual school basis and also summarized at the school district level. One of the summary measures reported at the school district level is "percent proficient or higher." This is an indication of the percentage of all schools in the district that achieved a score of proficient or better. These were downloaded from the MCAS website (`http://profiles.doe.mass.edu/state_report/mcas.aspx`) and filtered in Microsoft Excel to contain just the percent proficient or higher data for grade 10 English testing. These values were joined to the *School_Districts* shapefile and are represented as the attribute PA_pct. Now you'll analyze this data and make a map of it.

1 On the *School_Districts* layer properties *Symbology* tab, change the *Value Field* to ***PA_pct*** and click *Classify* to view a histogram of PA_pct values. In this dataset, the value zero (0) indicates "No data." Click the *Exclude* button and fill out the query by double-clicking *PA_pct*, clicking the equals sign (=), and typing ***0***. Your *Data Exclusion Properties* screen should look like the figure below.



Data Exclusion Properties

Query | Legend

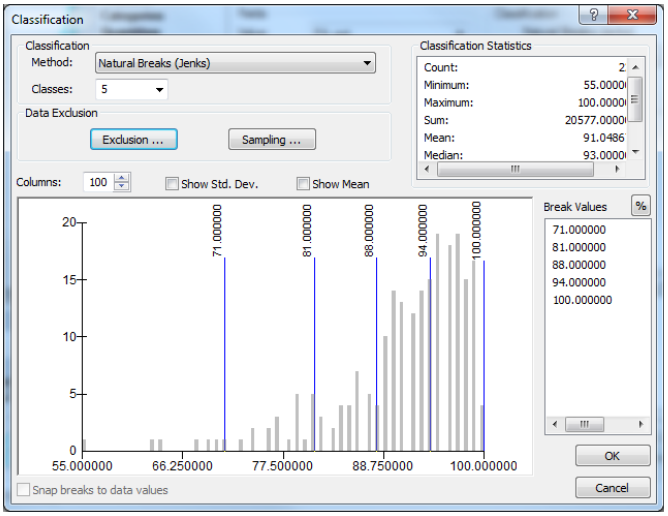
Exclude clause:

"FID"
"DISTNAME"
"PA_pct"

SELECT * FROM School_Districts WHERE:

```
"PA_pct" = 0
```

2 Click *OK*. Now examine the histogram. It should look like figure 1 below.



Classification

Classification Method: Natural Breaks (Jenks)

Classes: 5

Columns: 100 □ Show Std. Dev. □ Show Mean

Classification Statistics

Count:	2
Minimum:	55.0000
Maximum:	100.0000
Sum:	20577.0000
Mean:	91.0486
Median:	93.0000

Break Values

71.000000
81.000000
88.000000
94.000000
100.000000

□ Snap breaks to data values

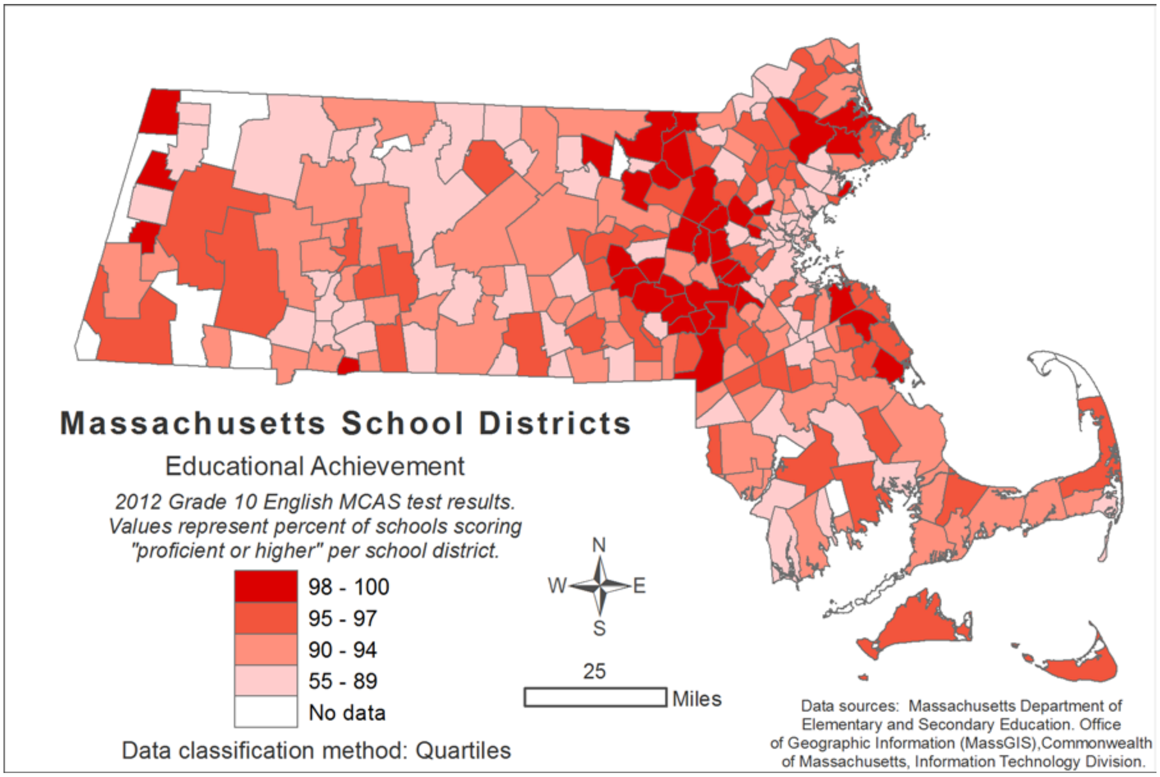
Figure 1: Histogram of MCAS grade 10 English percent proficient by school district.

Question 11: *Analyze the histogram. What is the total value range? How is the data distributed?*

3 Open the *School_Districts* attribute table and sort the *PA_pct* field in descending order.

Question 12: *What school districts had 100% testing of proficient or higher?*

4 Now make a choropleth map of educational achievement by school district. Here's an example of one in the figure below.


Massachusetts School Districts
Educational Achievement
2012 Grade 10 English MCAS test results.
Values represent percent of schools scoring
"proficient or higher" per school district.
98 - 100
95 - 97
90 - 94
55 - 89
No data
Data classification method: Quartiles
N
W
E
S
25
Miles
Data sources: Massachusetts Department of Elementary and Secondary Education. Office of Geographic Information (MassGIS),Commonwealth of Massachusetts, Information Technology Division.

Map 2. Educational achievement.

Question 13: *Using a visual comparison between maps 1 and 2, is the correlation of high income/high educational achievement immediately visible? What other correlations are apparent on the map?*

After visualizing correlations between these two variables, it's time to quantify any correlations using the numeric data and map them. A hypothesis to test with this exercise could be "school districts with a high amount of family income also have high-performing students." This could be based on the idea that a wealthier community leads to better education for students, which leads to higher test results. With this line of thinking, it would also be interesting to identify school districts that exhibit:

- Low family income yet still have high educational achievement
- High family income and low educational achievement
- Low family income and low educational achievement

Now you'll make a map showing districts that meet these criteria. You'll do this by using a very powerful tool: *Select By Attributes*. How to determine what is high and what is low? Arbitrarily choosing high and low cutoff values is one way to do this, but this method is very subjective, and values chosen could vary widely based on the person making the selection.

A different method to select high and low values from a dataset is to use *quantiles*. Quantiles are breaks that divide data into equal-size groupings. For example, grouping a dataset into four quantiles would result in four groups, each containing 25% of the data values. Using four quantile groupings is called *quartile*. Using five quantile groupings is called *quintile*.

So for "high" values, you'll use the top quartile for both median household income and educational achievement. From map 1, it is observed that the top quartile (aka "the top 25%") for median household income ranges from $91,143 to $153,748. From map 2, it is observed that the top quartile for percent proficient is from 98 to 100. Now you'll select school districts that meet these two criteria using *Select By Attributes*.

5 Click *Selection » Select By Attributes*.

6 For *Layer*, specify *School_Districts_BG_J*. Double-click the *"Avg_MHI"* text, then the greater than or equal to sign (>=), and type ***91143***. Click the *And* button, and then double-click the *"PA_pct"* field, the >= button, and type ***98***. Your expression should look like this:

```
"Avg_MHI" >= 91143 AND "PA_pct" >= 98
```

7 Click *OK*, and the school districts meeting these criteria will be selected.

8 Add a new Text field named ***Correlate*** to the attribute table of the *School_Districts_BG_J* layer. Right-click this field and click *Field Calculator*. Enter ***'High High'*** as the value. Now these selected districts are marked as containing high family income and high test scores in the *Correlate* field.

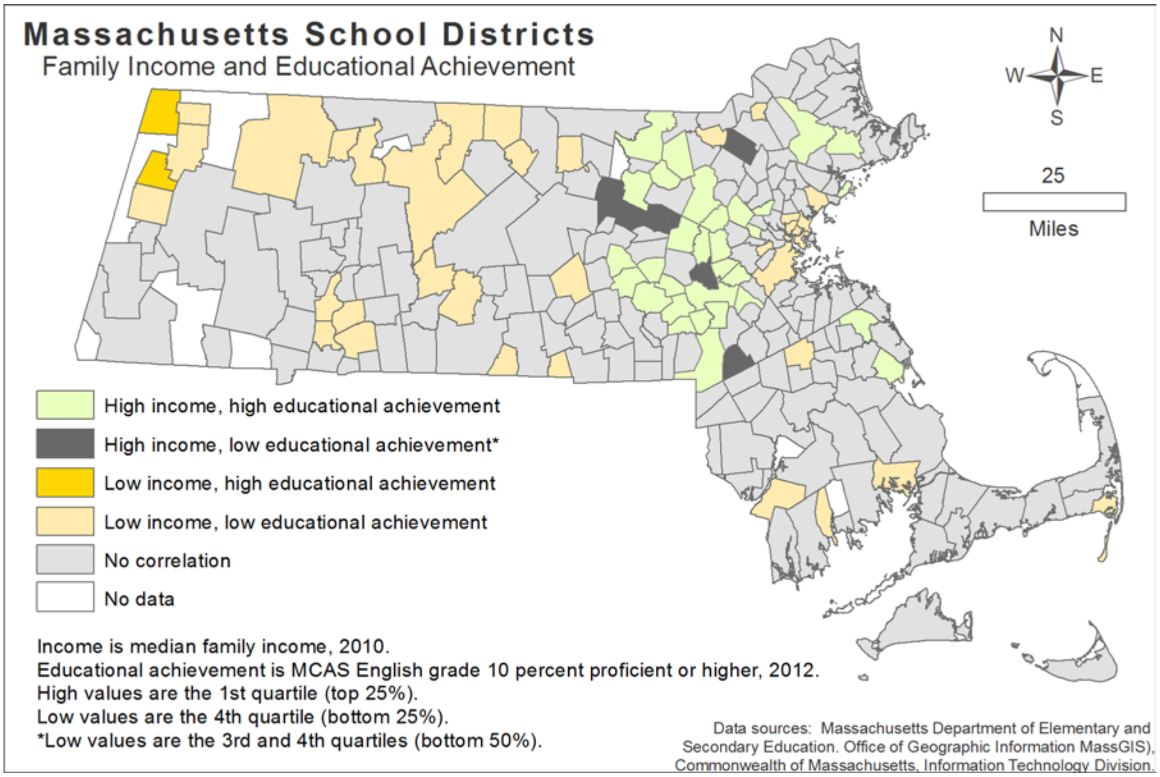
9 Use *Select By Attributes* again to select all districts in the LOWEST median family income quartile and in the HIGHEST educational achievement quartile. Calculate the *Correlate* field for these to ***'Low High'***.

10 Use *Select By Attributes* again to select all districts in the LOWEST median family income quartile and in the LOWEST educational achievement quartile. Calculate the *Correlate* field for these to ***'Low Low'***.

11 Use *Select By Attributes* again to select all districts in the HIGHEST median family income quartile and in the LOWEST educational achievement quartile. Calculate the *Correlate* field for these to ***'High Low'***.

12 Make a map that shows school districts with high/high, high/low, low/high, and low/low correlations. Do this by changing the symbology of *School_Districts_BG_J* to the *Show: Categories* symbology type and specifying *Correlate* as the *Value Field*. Click the *Add All Values* button to

bring in your four correlations as options to symbolize by. Add a legend, making sure to precisely describe what each color on the map means. Your map should look similar to the figure below.


Massachusetts School Districts
Family Income and Educational Achievement
N
W
E
S
25
Miles
High income, high educational achievement
High income, low educational achievement*
Low income, high educational achievement
Low income, low educational achievement
No correlation
No data
Income is median family income, 2010.
Educational achievement is MCAS English grade 10 percent proficient or higher, 2012.
High values are the 1st quartile (top 25%).
Low values are the 4th quartile (bottom 25%).
*Low values are the 3rd and 4th quartiles (bottom 50%).
Data sources: Massachusetts Department of Elementary and Secondary Education. Office of Geographic Information MassGIS), Commonwealth of Massachusetts, Information Technology Division.

Map 3. Median family income and educational achievement correlation map.

Notice on this map that the "High income, low educational achievement" category uses the third and fourth quartiles instead of just the fourth quartile. This change in classification was made because no districts were selected when choosing the fourth quartile in educational achievement and the first quartile in median family income. So for this selection, the low educational attainment range was expanded. This is a decision by the cartographer that significantly affects the map, and thus is duly noted there.

Question 14: *Use the Identify tool to determine which school districts have low income and high educational achievement. What districts are these? Do these school districts exhibit any spatial pattern?*

Question 15: *Which district has low income and high educational achievement?*

Submit your work

Submit the following to your instructor:

- Map 1: Median family income by school district
- Map 2: Educational achievement
- Map 3: Median family income and educational achievement correlation map

Credits

Sources of supplied data

Block_Groups, courtesy of Esri Business Analyst 2012.

School_Districts, courtesy of Office of Geographic Information (MassGIS), Commonwealth of Massachusetts, Information Technology Division.

Town_Boundaries, courtesy of Office of Geographic Information (MassGIS), Commonwealth of Massachusetts, Information Technology Division.

Instructor resources

Context for the lab

This *SpatiaLAB* is written primarily for undergraduates studying GIS, cartography, or education.

It uses spatial analysis software and publicly available geographic data to explore the relationship between family income and educational performance. Along the way, students gain an understanding regarding the process used to prepare the data in order to perform analysis on how two different phenomena may or may not be connected. While this exercise focuses on analyzing family income and educational performance in Massachusetts, the process learned will be applicable to analyzing two variables occurring in other locations.

The lab shows how to use geographic information to create choropleth and correlation maps. It is intended to promote thinking about how maps can be symbolized to evoke certain interpretations of various geographic phenomena and how data can be used to reveal correlations.

The learning objectives of this exercise are as follows:

- To create aggregate statistics and visualize them geographically
- To analyze a statistic and determine how to best classify it for choropleth map display
- To recognize and analyze spatial patterns
- To realize the utility of analyzing things geographically

Instructors may engage students in discussion related to (1) additional criteria that may need to be considered to make determining correlations more informative; 2) expanding the MCAS datasets to include additional years, grades, and test types; and (3) alternative mapping techniques to best communicate these phenomena and correlations.

Using a spatial approach, you will find GIS to be one of the most useful tools for performing mapping and analysis tasks to analyze any two variables that can be mapped.

This lab uses town boundary, school district, block group, and MCAS data.

Students are asked to answer 15 questions, perform GIS analyses, and make three maps.

Analysis and visualization tools

ArcGIS 9 or 10 is required to complete this lab.

Answers to questions

Question 1: *How many high school districts are there in Massachusetts?*

Answer: 228.

Question 2: *Describe the spatial relationship between towns and school districts. Which one contains larger areas on average? Are the boundary lines between towns and school districts coincident?*

Answer: School districts are larger than towns on average. Many school district boundaries are coincident with town boundaries, but in some cases multiple towns are aggregated into one school district.

Question 3: *What are census block groups? Is there a small variation or large variation in the size of a census block group? Use research from the web and cite your information sources.*

Answer: A geographic unit defined by the US Census Bureau. It is the smallest unit for which the census bureau publishes sample data. Block groups exhibit a wide variation in size.

Question 4: *The most populous block group is in what town/city?*

Answer: Amherst.

Question 5: *Name two of the towns/cities that have a block group with zero population. Using the Imagery Basemap, describe what you see in these zero-population block groups.*

Answer: Any of the following is OK: Chicopee, Boston, Stow, Winthrop, Somerville, Freetown. The Imagery basemap shows that in block groups with zero population, these are often parks, beaches, forests, airports, or other areas with no residences.

Question 6: *What are the mean, minimum, and maximum median household incomes for school districts in Massachusetts? What is the variation between the min. and max. values? Is this disparity (or lack of) surprising?*

Answer: Mean: 76,013; minimum: 36,403; maximum: 153,747. The difference is 117,344.

Question 7: *What classification technique might be the best for identifying outliers in the dataset (i.e., either extremely high or extremely low data values)?*

Answer: Standard deviation seems to isolate the low outliers. For the high outliers, a manual classification could be used.

Question 8: *What effect does increasing the number of classes have on the map? What effect does decreasing the number of classes have on the map?*

Answer: Increasing the number of classes makes the map smoother (but hard to interpret if too many color shades); decreasing makes class differences stand out more on the map but generalizes more.

Question 9: *What is a "diverging color scheme"? With what kind of data would this scheme be used?*

Answer: A diverging color scheme involves one color decreasing in value, and then a neutral color, and then another color increasing in value, all in the same sequence. It is often used to portray quantitative data that is divergent around a critical midpoint value. A type of data may be a map showing areas of population increase, remaining the same, and then a decrease.

Question 10: *Examine the map. What is the spatial pattern of family income in Massachusetts?*

Answer: High in eastern Massachusetts and a mix of low and medium in central and western Massachusetts.

Question 11: *Analyze the histogram. What is the total value range? How is the data distributed?*

Answer: 55.0 – 100.0.

Question 12: *What school districts had 100% testing of proficient or higher?*

Answer: Harvard, Winchester, Dover-Sherborn, Northboro-Southboro.

Question 13: *Using a visual comparison between maps 1 and 2, is the correlation of high income/high educational achievement immediately visible? What other correlations are apparent on the map?*

Answer: Yes, in the suburbs of Massachusetts just west of Boston. No other correlations are visible.

Question 14: *Use the Identify tool to determine which school districts have low income and high educational achievement. What districts are these? Do these school districts exhibit any spatial pattern?*

Answer: Nashoba, Tewksbury, Foxborough, Natick, Maynard. Yes, they are all roughly at the same longitude (aligned north to south — suburban Boston).

Question 15: *Which district has low income and high educational achievement?*

Answer: Mount Greylock.

Data information

The dataset MA_Education.zip contains the town boundary, school district, and block group data.

Data sources

Town Boundaries, School Districts: Office of Geographic Information (MassGIS), Commonwealth of Massachusetts, Information Technology Division.

Specific URLs for this data:

MCAS scores: Massachusetts Department of Elementary and Secondary Education. `http://www.doe.mass.edu/mcas/results.html` (accessed October 2012).

Block group boundaries: Esri Business Analyst 2012.