

11.520: A Workshop on Geographic Information Systems

11.188: Urban Planning and Social Science Laboratory

Lab Exercise 3: Database Operations in ArcMap

Due Lab 4

Overview

In this exercise, you will use the Cambridge housing sales and census block group data to explore ArcMap's database query capabilities involving query selection, tabular joins, and spatial joins. The purpose of this part of the lab exercise is to introduce ArcMap's ability to perform analyses that depend upon data queries and the juxtaposition and manipulation of spatial features. Much of the analytic power of a GIS comes from its capacity to compute and then manipulate and visualize the geospatial relationships among selected events and locations. We'll illustrate a few basic spatial analysis techniques using data sets from the previous ArcMap lab exercises --namely, the 1989 housing sales data for Cambridge and 1990 Census data for Cambridge block groups. Subsequent labs will cover additional capabilities. Also this exercise provides opportunities to set appropriate data coordinate system, which was covered in the lecture. In this particular lab you will turn in your answers by filling out a form. As you work on the exercise, please print out (or edit) the answer form and write in your answers.

(1) Merging, editing, and saving tables: To do this lab, we'll have to learn a few more ArcMap techniques for merging and saving tables. There are hundreds of census variables and any Cambridge blockgroup map is likely to include only a small subset of them. To use other census data, we'll have to load these data into ArcMap and "join" them with the attribute table of the Cambridge blockgroup theme using a common geographic reference variable (such as the state-county-tract-blockgroup identifier, **Stcntrbg**). We may also need to compute new fields that normalize or otherwise combine multiple columns of the original data. But the official class datasets are "read-only" and ArcMap requires that you have write access to attribute tables in order to create and calculate new fields. So we'll also need to learn how to save local "writeable" copies of (extracts of) our class datasets.

(2) Spatial Joins: Suppose we'd like to compare the sales prices of Cambridge homes with the socio-economic characteristics of their neighborhoods. For example, we might want to know whether the high-priced sales tended to be in neighborhoods with highly educated adults. By drawing a pin-map of high-priced 1989 sales locations on top of a Cambridge block group map, we may be able to see a pattern. The block group map can be thematically shaded by educational attainment levels. With this map, we can 'see' whether high/low-priced sales cluster in neighborhoods with, for example, more (or less) educational attainment. But the pattern may be misleading or hard to interpret and we may want to quantify the relationship to measure the degree of association and to tag the sales data with the some of the characteristics of their neighborhoods. Much of the benefit of using GIS software depends on its tools for cross-referencing and reinterpreting data based on spatial location. We'll defer until later labs the more general and advanced tools for tagging data based on spatial proximity and we'll focus in this lab on simpler spatial joins and buffer creation.

(3) Setting coordinate system: So far, the coordinate system of the data has already been preset for your use. However, in typical settings, you may need to specify or change the coordinate systems associated with GIS datasets. In this exercise, you will learn how to apply a suitable coordinate system to your datasets.

I. Setting up your Work Environment


Follow the usual routine to set up your work environment:

- Log onto Athena PC
- Attach drive M: by using the DOS command "attach -Dm 11.520".
- Launch a web browser and open the web page for the current lab.
- Start ArcMap

II. Examining Attribute Data

As in previous labs, we will use Cambridge data for this exercise. In ArcMap, start a new map document and add **Cambbgrp.shp** in **M:\www\labs\lab3** directory to the data frame. Open the data frame property window by doubling clicking data frame name **layers**, rename it to "Lab Exercise 3" and set the "Display Units" to miles. Remember: you need to set the map and Display units in every new view you create in order for ArcMap to interpret the coordinates properly and generate correct scale bars and distance measurements. Make a habit of doing this immediately after adding the first layer to a new map document.

A. Simple Queries

In an earlier lab, we've already learned how to use the *Info button*, , to provide attribute information about particular spatial features. We have also learned how to open an attribute table of the geographical layer.

Now we will experiment with querying the data. Select **"Select by Attributes"** under the **"Selection"** menu. On the **"Select by Attribute"** window, make sure choose **Cambbgrp** for the "layer" window. In this case, we don't need to worry about that because we have only one layer. Otherwise, we have to pick up the correct layer from the drop-down list. In the method box, keep the default one "create a new selection".

Now you can type your query manually into the box in the bottom left corner of the window, or you can use the tools at the top of the window.

Let's find the block groups with median household income in excess of \$50,000 per year. In the "Fields" list, double-click on **"Med_hh_inc"**. Then single-click on the ">" button. Next, you can type in 50000 or scroll down "Values" list until you can double-click on 50000. The query entry window should now show **"Med_hh_inc"> 50000**

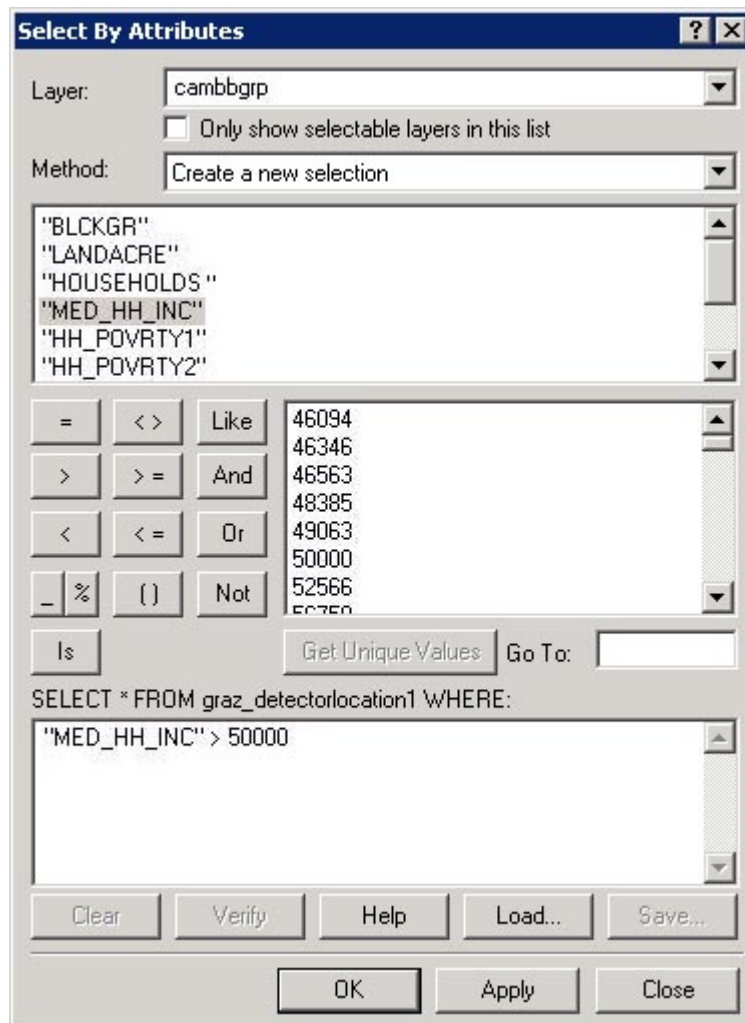


Fig. 1. Query Window

The window finally looks like the Fig. 1. Click "Apply" button to run the query and 'select' these block groups. Notice that all block groups where the median household

income is greater than \$50,000 are now highlighted in the data display area. If you open the attribute table, you can find out the associated records are also highlighted. You may need to scroll up or down in the attributes window to see the highlighted records. If you click the **Selected** button, which is at the bottom of the attributes window, only the selected records will be visible. Next to the **Selected** button, you can find "9 out of 94 are selected".

B. Statistics and Selected Sets

Now close the query window. Make sure the selected records are still highlighted. Select **Statistics** from the **Selection** menu. You should see a selection statistics window. Choose **Med_hh_inc** in the field drop down list (Fig. 2). You should see a window listing some common statistical measures describing the 9 selected records of the **Med_hh_inc** field. It also generates a frequency distribution diagram.

Now close the selection statistics window and return to ArcMap window. Swap the selected and unselected block groups in the following way: 1. Open the attribute table; 2. Click **Options**; 3. On the pop-up menu, select **Switch selection**.

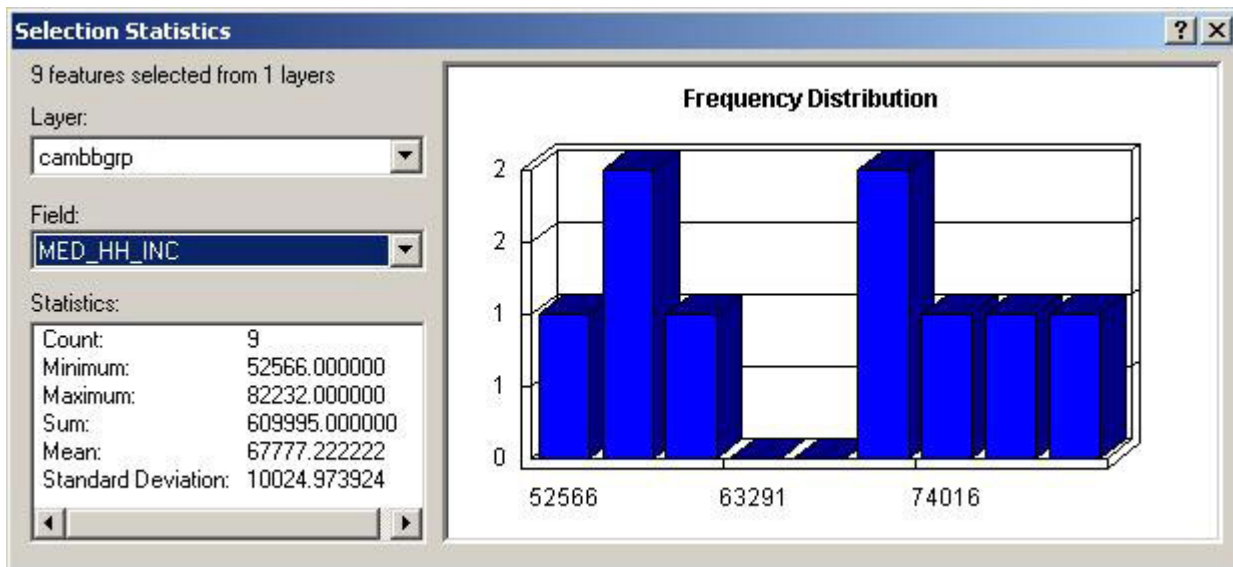


Fig. 2. Statistics Window

Fig. 3. Switch Selection

Select **Statistics** from the selection menu again to summarize these 85 lower income block groups. **Write the mean and standard deviation of these 85 block groups on your lab assignment.** (There is more than one way to access the selection statistics. With the attribute table open, try right-clicking on the column heading of the field for which you want statistics...)

Note from the statistics for these 85 block groups that at least one block group had a median household income of zero. Let's look at these block groups using the **select by attributes** function. We find that there are with zero income. Two of these block groups

are near Tech Square and one of them is near Harvard Square (where students in dorms count as people but not households). Let's exclude these block groups from consideration and recompute the statistics for the 82 block groups with a median income of \$50,000 or less. Specify the selection criterion as "**MED_HH_INC**" > 0 and "**MED_HH_INC**" <=50000. Then use the *Selection > Statistics* menu to compute the mean and standard deviation for those block groups with median household income greater than zero and less than or equal to \$50,000. **Record these two numbers on your lab assignment.** Since we consider the census data to be unreliable for the three block groups with '0' income, we may wish to exclude them from consideration entirely. One way to do this is to set the layer properties for **Cambbgrp**. Double click on the layer's name to bring up the layer property window. Under the tab "definition query", type in "**MED_HH_INC**" > 0. Click OK. Notice that the **Cambbgrp** layer no longer shades the three block groups with zero income. They have been excluded from the layer and they are no longer shaded, but are 'donut holes' in the Cambridge map. The data are still there on disk but ArcMap is ignoring the three block groups that don't meet the selection criteria specified in the layer definition criterion. For the remainder of the exercise we will exclude the zero-population block groups from our analysis.

C. Saving and Editing Data Tables

Suppose we wanted to generate a thematic map of the percentage of people (aged 25 and older) who have less than a high school education. The **EDU** variables in the Attribute table of **Cambbgrp** table provides the required information. They indicate the 1990 census counts of persons (aged 25 years and over) with various degrees of education as in Table 1.


Table 1. EDUCATIONAL ATTAINMENT (Universe: Persons 25 years and over)	
Column	Description
EDUTOTAL	Total Persons 25 years and over
EDU1	Less than 9th grade
EDU2	9th to 12th grade, no diploma
EDU3	High school grad (includes equivalency)
EDU4	Some college, no degree
EDU5	Associate's degree
EDU6	Bachelor's degree
EDU7	Graduate or professional degree

From the listing, we see that the desired percentage equals

$$100 * (\text{EDU1} + \text{EDU2}) / \text{EDUTOTAL}$$

The tables we are using are read-only so we can't edit the tables in order to add a new column that computed this percentage. To overcome this problem, we'd like to create (and edit) a local copy of the table. But there are several dozen columns in the table and we are concerned only with the geographic identifiers and the **EDU** fields.

Open the **layer properties** window, go to the **Fields** Tab. Turn the 'visible' off for all but the three needed **EDU** fields and the **Stcntrbg** blockgroup identifier. To turn off a field, first select the field, then uncheck the visible box. When you click 'OK' the table will show only those four fields. Open the attribute table again to confirm this behavior.

Now you need to export the table. Click on the **Option** Button, which is on the right bottom of the attribute table window. On the pop-up menu, pick up **Export**. Save the table to your working directory. If you have not registered your working directory with **ArcCatalog**, you need to use the **connect to folder**  button to register it first. Save the table as a dBase-formatted table. Call the table **education.dbf**. The *.dbf suffix will be a handy reminder of the data format for the table and the dBase format is quite portable and easily read by Excel and many other packages.

When asked whether you want to add the table to the current map, click Yes. Now this newly saved-to-disk **education.dbf** table will be included in your map document file. However, you may not be able to see it in the data frame if the bottom tab is set to **Display** instead of **Source**. Change the view of the data frame from **Display** to **Source**; and the **education.dbf** table will be listed. If you clicked the wrong button when asked to add the table to the map file, it won't be listed, but you can now add it in the same way as you add any other map layer or attribute table.

Since you now have write access to your own **education.dbf** table, you will now be able to add a new column for computing your low-education percentages. Right click the table name in the data frame. On the pop-up menu, select **Open Attribute Table**. On the new window, click on options, and on the pop-up menu, select **Add field**. Call the new field **p_lowed** and change the data type to float. Click OK, and you will see a new field appear in the window. Next, right click on the new column name and on the pop-up menu, select **Calculate Values**. Click "yes" when a warning message comes up (about not being able to undo the editing). Then set

p_lowed = (this top line should appear by default above the box with the cursor and you should not retype it)

100 * ([EDU1] + [EDU2]) / [EDUTOTAL]

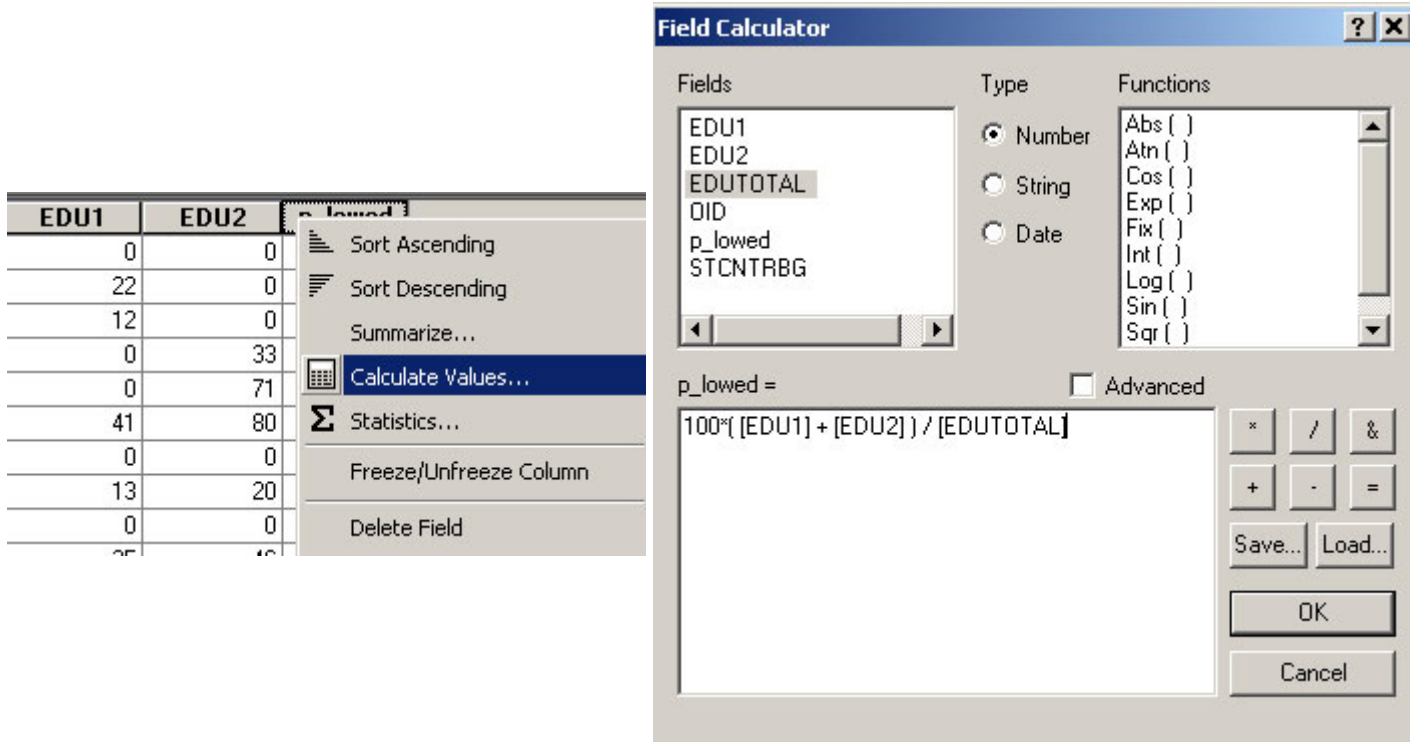


Fig. 3. Calculate Values

Sort the table (in descending order) by **p_lowed**. Right click on the column name and select **Sort Descending**. The highest **p_lowed** value is 55.5%. What is the second highest **p_lowed** percentage? How many block groups have more than 50% of their adults lacking a high school diploma? *Write the values on your lab assignment.*

D. Merging Data Tables (via Joins and Relates)

Next, we'd like to generate a thematic map of the newly computed **p_lowed**. In order to do this, we need to 'join' the new table with our percentages to the **Attributes of Cambbgrp** table that is linked to the map. The **STCNTRBG** field is a unique identifier that appears in both tables and can help us cross-reference the two tables. To link these two tables together, first open the **layer properties** window of the **Cambbgrp** layer (double click the name of the layer). Go to the **Joins & Relates** tab. In the **Join** frame, click the **Add** button. The "Join Data" window appears. It allows you to specify the criterion. First, select **Strcntrbg** as the field of the GIS layer that the join will be based on. Second, select **Education** as the table (**education.dbf**) that will be joined to the GIS layer. Since we have already added the table into the project, we can pick it up from the drop-down list. Otherwise, we need to add it from the disk. Thirdly, select **Strcntrbg** as the fieldname on which to base the join. This tells ArcMap to join the **education** table to the GIS layer **Cambbgrp**.

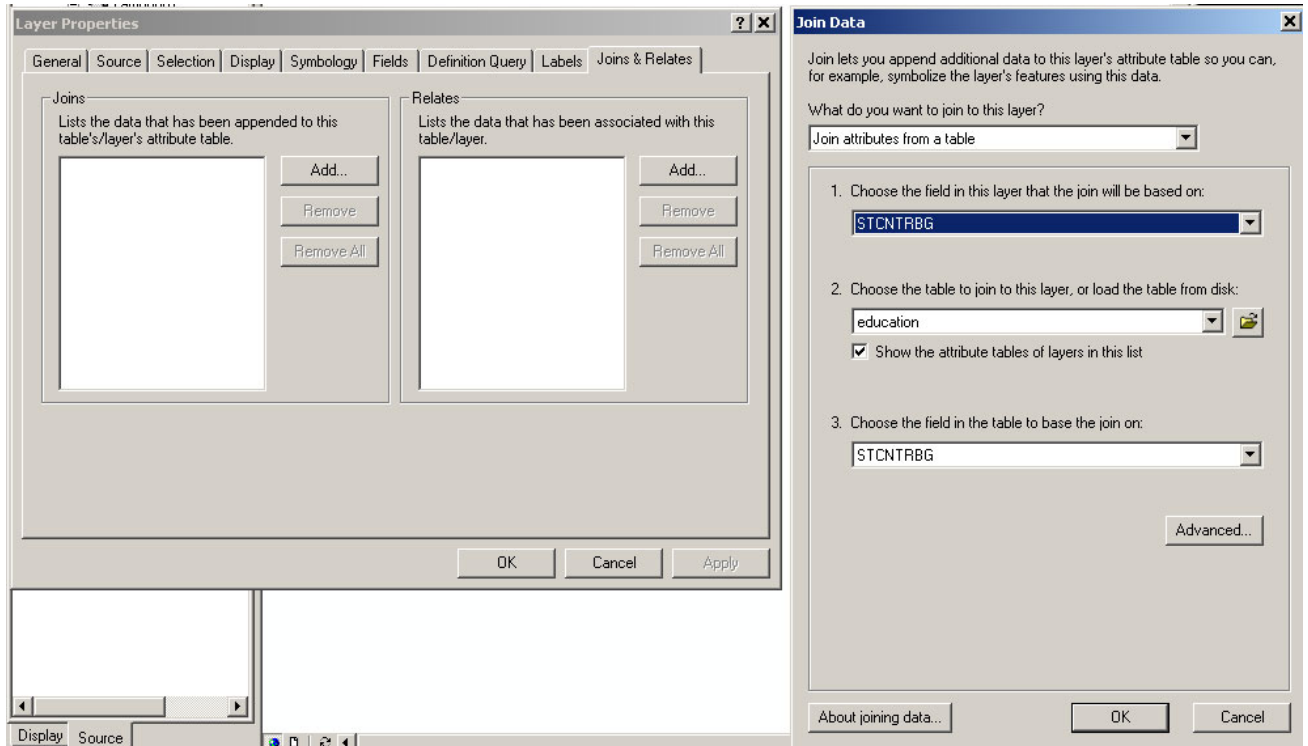


Fig. 4. Join


Apply the join and go back to the ArcMap window. Open the attribute table of **Cambbgrp**. We now find the columns of the education tables appended to the right side of the attributes of **Cambbgrp** table. Now, we are ready to shade a thematic map of the newly computed **p_lowed** field. Open the layer definition window. Go to Symbology the. Classify your **p_lowed** field using 5 quantiles categories and eyeball the results. To save time, we won't ask you to print out this map. Just pick a readable graduated color scheme and take a little time to examine the thematic map: Do the patterns you see match your impression of the socio-economic patterns in Cambridge?

Please note that the table 'joins' are one-directional. If you join the tables in the reverse order, the **Attributes of Cambbgrp** table will disappear and you won't be able to map the selections from the expanded **education** table. Finally, note that these Joins are temporary and do not affect the tables that are stored on disk.

III. Spatial Joins and Buffering

Thus far all our query examples have focused on calculations and selection criteria that are done directly on attribute tables. We would also like to be able to perform queries that depend upon the spatial relationships of the spatial features that are related to the rows in the table.

A. Simple Graphical Selection

We are already familiar with the simplest graphical selection tools . Suppose we wanted to use the map to examine a few Cambridge block groups near MIT. This is easily done using the graphical selection tools and *Statistics* command from the Selection menu, which we have learned in previous labs and in Part I above. In the data display area, select the two block groups along the southern-most edge of Cambridge that contain most of the MIT campus. Now use the *Selection > Statistics* menu item to determine the mean **population** for those two block groups. It should be 2535.5.

B. Spatial Joins Using a Quick Form of Point-in-Polygon Analysis

Next we'll exercise the "spatial join" capabilities of ArcMap to see whether lower-priced housing tends to be in neighborhoods with relatively low income and low levels of education. We've already mapped the low education percentages, **p_lowed**, for Cambridge block groups, and the **sales89** table (in **m:\data**) contains the location of all (1-4 family) residential homes in Cambridge that sold during 1989. These sales data come from a Banker and Tradesman Real Estate Transfer Database, for 1987-1989 (data that Anne Kinsella Thompson acquired for use in her MCP thesis). We can address our question about housing value, income, and education, if we can observe and summarize the extent to which the low-priced housing falls within those block groups with high **p_lowed** values (or low **med_hh_inc** values, if we map the median income instead).

Add the **sales89** coverage as a layer in your ArcMap window. Open the **layer properties** window and go to the **definition query** tab and create query definition to include in your layer *only* those sales with a **realprice** less than \$150,000. You should find that 29 of the 222 sales meet this criteria. Now select areas of the city (block groups) where these sales occurred. Go to menu *Selection > Select by location*. Specify the criterion as in Figure 3.

Click the Apply button. Every block group that contains one or more of the low-priced sales will be highlighted. You've just done a basic **"point-in-polygon"** query to find a set of polygon features (block groups) which contain a set of point features (the low-priced sales).

Analyzing the results:

- Select **Statistics** from the **Selection** menu
- Check the Mean value for **med_hh_inc** across these 21 block groups. It should be \$30,070.
- Without closing the statistic window, right click on the layer name **cambbgrp**. On the pop-up menu, go to **Selection->Switch Selection**.

- The statistics change automatically.
- Check the mean value for **med_hh_inc** across the other 70 block groups (the ones with no sales under \$150,000). It should be \$38,454.

The lower priced sales do appear to occur in block groups with somewhat lower income. Next, do the same queries for the **p_lowed** values that you computed earlier. What is the (unweighted) mean and standard deviation of **p_lowed** for the 21 block groups that contained all the low-priced homes (**realprice** < \$150,000)? What about the other 70 block groups? **Write the values on your lab assignment.**

These point-in-polygon queries are useful for quick exploration of the data but the summary statistics are only that -- a summary of the patterns that result. As we might suspect, the general trend suggests that low-priced housing tends to be in lower-income,

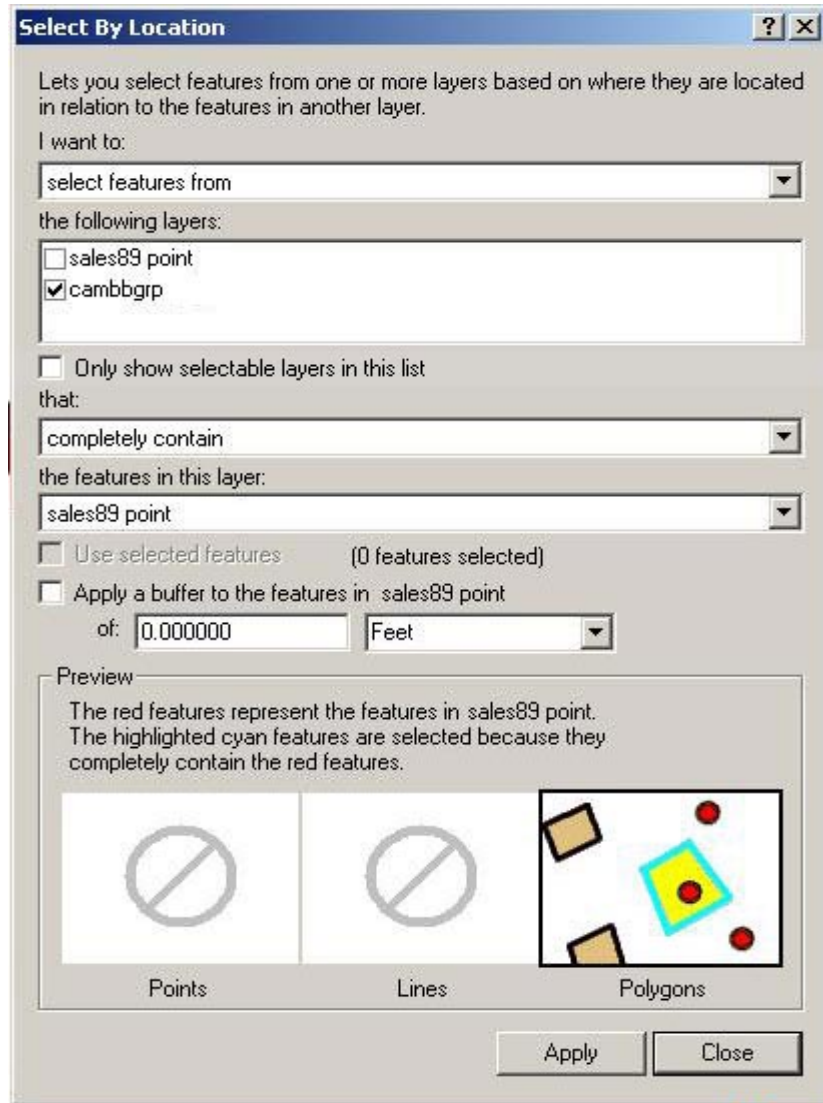


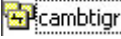
Fig. 5. Select by location

lower-education neighborhoods. But there is quite a bit of variability and the tools we've used so far don't let us move the **Attributes of Cambbgr** data from the census table over to the appropriate **sales89** rows. Doing that would permit us to examine the patterns more closely. In later labs, we'll look at other tools (in the ArcGIS Toolbox) that let us tag the **sales89** table with the census data for the blockgroup that contains the sale.

C. Simple Buffering

The "spatial join" in the previous section involved asking which set of spatial features (sales) were completely contained within another set (block groups). Another simple "spatial join" operation is to determine which spatial features are **close to** other spatial features. Buffering tools are one way to do this.

Suppose we want to check whether the lower-priced housing tends to be closer to the major roads. Let's use ArcMap's buffering tools to create a buffer around Mass Ave and see whether the sales in the buffer are relatively high or low priced.

Add **cambtigr** coverage  to your ArcMap project. Choose to add the 'arc' features from this coverage. The **cambtigr** coverage is stored in **M:\data**. This is the Cambridge street coverage that we used in an earlier lab. Let's add a second copy of this layer to our "Lab Exercise 3" data frame. (We're doing this so we can limit one copy of the road theme to a single road -- Mass. Ave.) Right click on the layer name **camtigr** and select 'copy'. Right click on the data frame name "Lab Exercise 3" and select 'paste layer'. Open the layer property window of one of the TIGER layers. Under the tab **general**, change the layer name to **Mass Ave**. Go to the **Definition Query** tab, and use the query builder to limit it to only those road segments with **FNAME = Massachusetts**. (There is one street segment on the Mass Ave bridge that has **FNAME = 'Massachusetts Ave'**! Don't bother including that link.) Beware of two issues when making this selection: (1) the field name shows up as 'fname' in the 'definition query' window, but is listed as 'fname' if you use the 'identify' cursor to click on the street links display in the map window. That's because column names are allowed to have aliases - take a look at the 'fields' tab in the layer properties window and you'll see both names! (2) Also, note that some Mass Ave street segments are not selected. That's because the **FNAME** for those segments are listed as 'State Hwy 2A' rather than 'Massachusetts'. This multiple-name issue could be a problem. In this case, the 55 selected Mass Ave street segments are sufficient to create a buffer that will enclose the others so we won't need to do extra work to identify those Mass Ave segments that list the route number instead of the street name. ArcGIS has some more elaborate database table schemas to handle such naming and route numbering issues but we won't get into that level of complexity in this exercise.

Next we will use the 'Buffer' tool in the ArcToolbox to create a buffer. We want to buffer the 55 features of our **Mass_Ave** theme by 0.5 miles. (Make sure you have already set the **Data Frame Properties** so that maps units are in meters and display units are in **miles**.)

Open the **ArcToolbox** window from the **Window** menu. Then choose the **Buffer** option from the '**Proximity**' listing under '**Analysis Tools**'. Choose **Mass Ave** for the input

feature and enter a path and shapefile name for the output feature that will store it. In the 'Distance' portion of the buffering window, set the linear unit to be 0.5 and be sure the units are set to miles. Accept the defaults for the 'side type' and 'end type' (regarding how the shape of the buffer is computed) and change the 'dissolve' setting to 'all' so that overlapping buffers around each street segment are merged. When you've adjusted all these settings, click "Finish" and wait for ArcMap to do the computations. You should

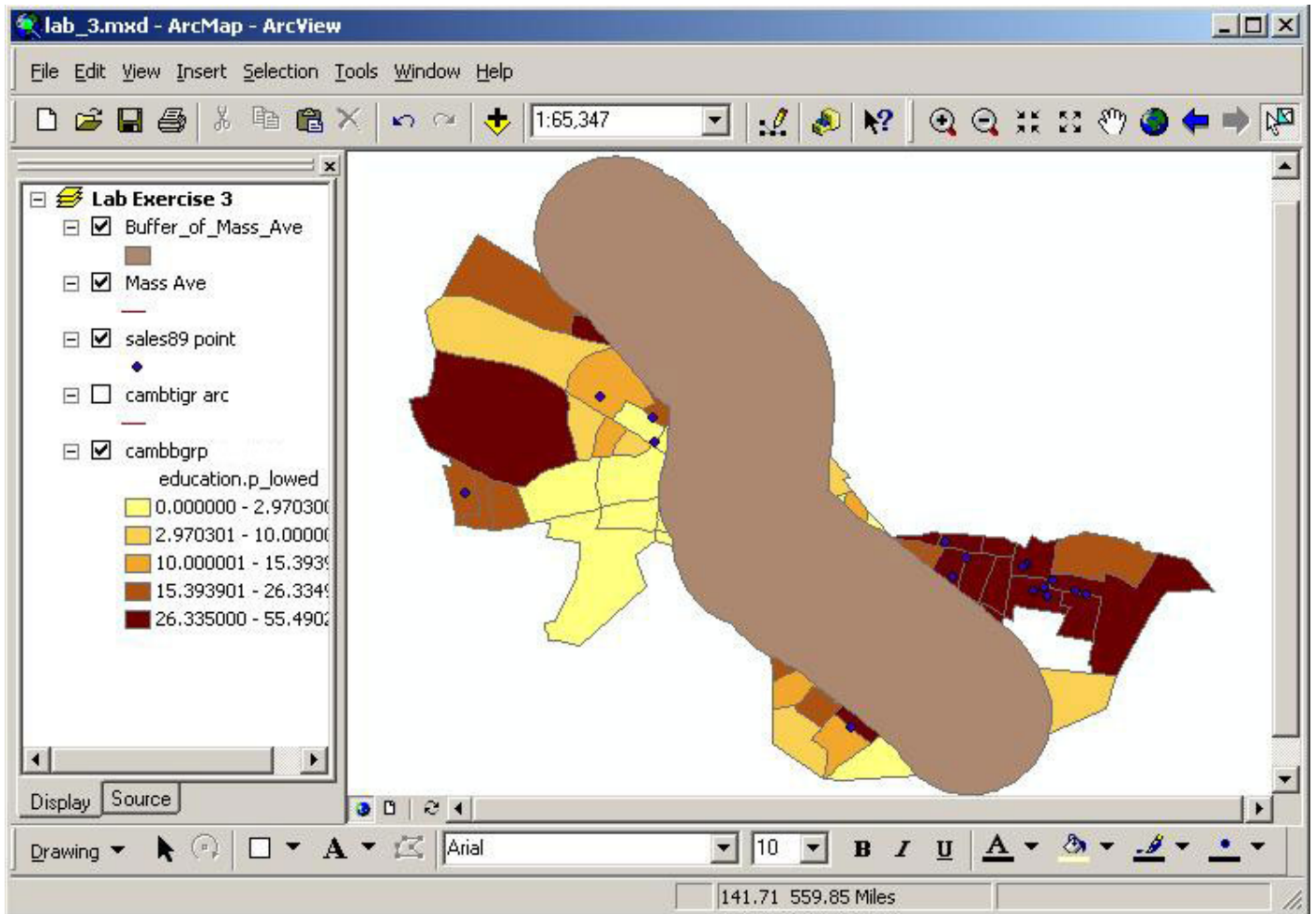


Fig. 6. Massachusetts Avenue Buffer

end up with a curved sausage-shaped area something like the one in Fig. 4.

Since this newly created buffer is a shapefile, we can use it just like any other theme. In particular, we can do an exercise similar to "point-in-polygon" example we just did in order to determine which 1989 housing sales are located within the new buffer. Highlight the **sales89** theme (with **all** the sales not just the low-priced ones) and choose **Selection > Select by Location**. Select all the **sales89** cases that **intersect** the **Buffer of Mass Ave** layer. (We used "completely contain" before and "intersect" this time. Do you understand why?) You should find that 140 of the 222 sales are located within the half-mile buffer. What are the mean and standard deviation of sale prices for the sales within the buffer the buffer? *Write the values on your lab assignment.*

The sales prices in and out of the buffer aren't all that different (compared with their standard deviation). Also, some parts of the buffer falls outside Cambridge and we don't know about home sales prices in those areas. Before reaching any conclusions, we would want to do further analysis. But, this quick tour of spatial selection is enough for today. In subsequent labs, we'll examine lots more of the spatial analysis capabilities of ArcMap.

IV. Coordinate system

In this part, you will simply apply a suitable coordinate system to a Middlesex county city boundaries layer for a map you will create. Suppose you would like to create a map layout shown below. As you can see, there is a small map in the left bottom corner. It shows cities in Middlesex County, Massachusetts and red area is Cambridge.

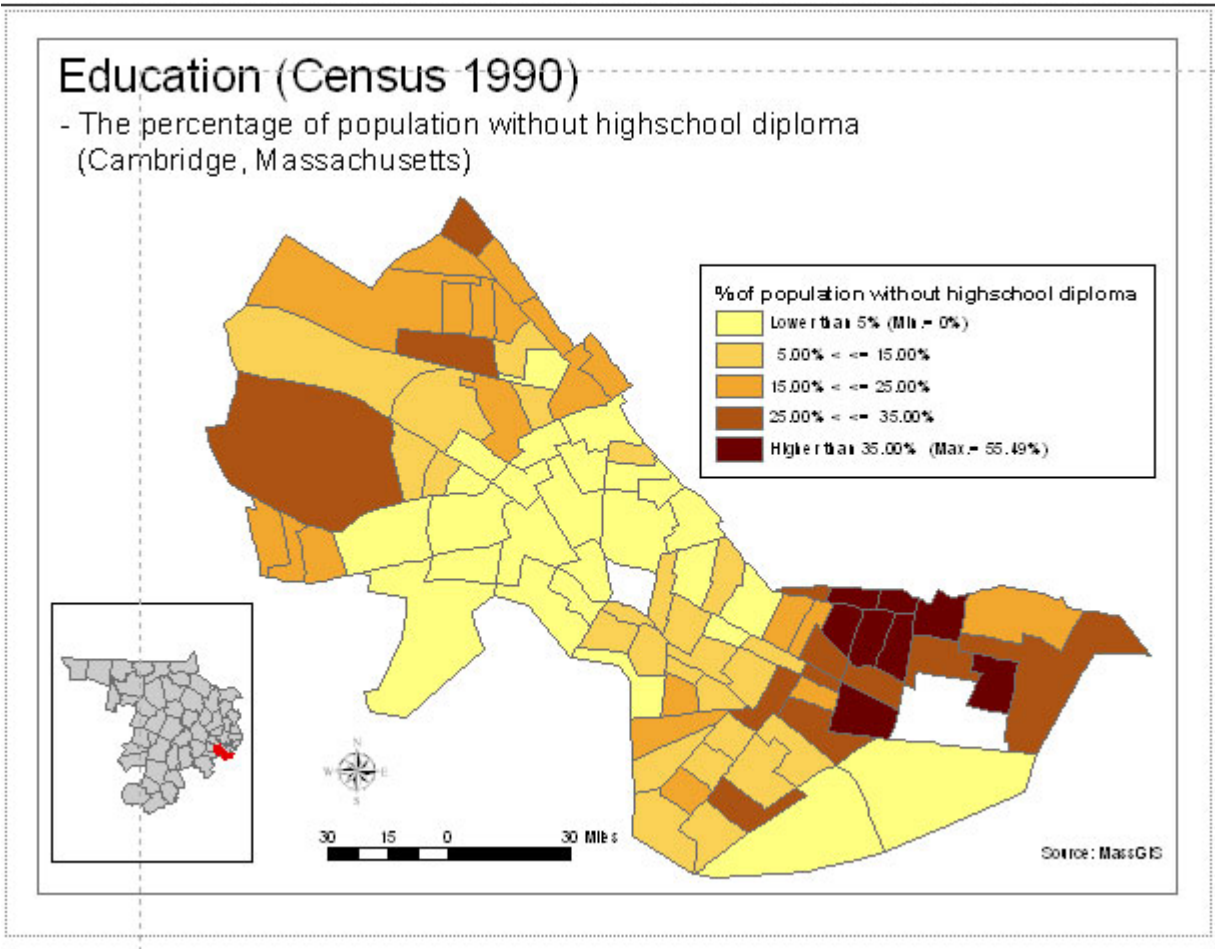


Fig. 7. Map layout

To create this map, you have to create a new dataframe as you did in lab2. Rename the new dataframe to **legend map** and copy **Cambbgrp** layer and paste into the **legend map** dataframe. Add a new layer **middle_county** to the **legend map** dataframe from **M:\data\lab3_county**. (The map layer shows all the towns within Middlesex County in Massachusetts). When you add the shapefile, you will get a warning message something like this:

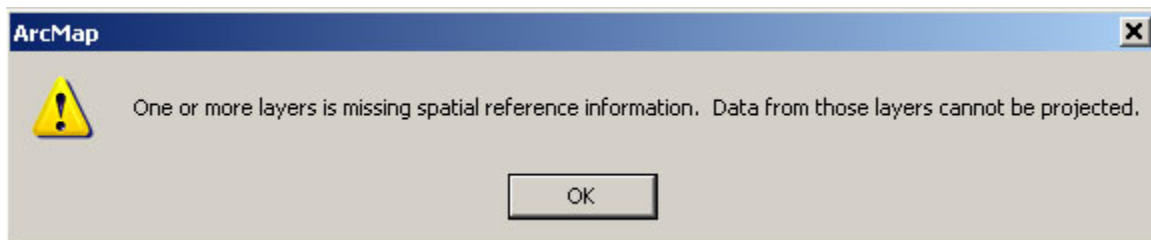


Fig. 8. Warning message

This message notifies you that the spatial reference information is missing from the **county** layer and may cause problems for further use. In some cases, you may not get an

error message but it does not mean that it is O.K. Let's check it out. Open the **middle_county** layer's **layer properties** window and select the **source** tab. It shows that the coordinate systems of this layer is "**Undefined**". Let's fix this. Close the layer properties window and remove **county** layer from the **legend map** dataframe. You cannot change the coordinate system of a layer while it is open in ArcMap. Now launch the ArcCatalog (**Start > All Programs > ArcGIS > ArcCatalog**).

1) Copy and paste the **middle_county** layer from **M:\www03\labs\lab3** into your drive. When you copy and paste the layer, please use ArcCatalog. A GIS layer consists of several files, such as .shp, .shx, .dbf, .sbn, .sbx, and you may miss necessary files when you use windows explore to copy the layer. Click right mouse button while put the cursor on the name of the layer. Drop-down menu will show up. Click **properties** from the list. Then **Shapefile Properties** window will show up. (Note that, in this graphic, the shapefile is called 'county' rather than 'middle_county'.)

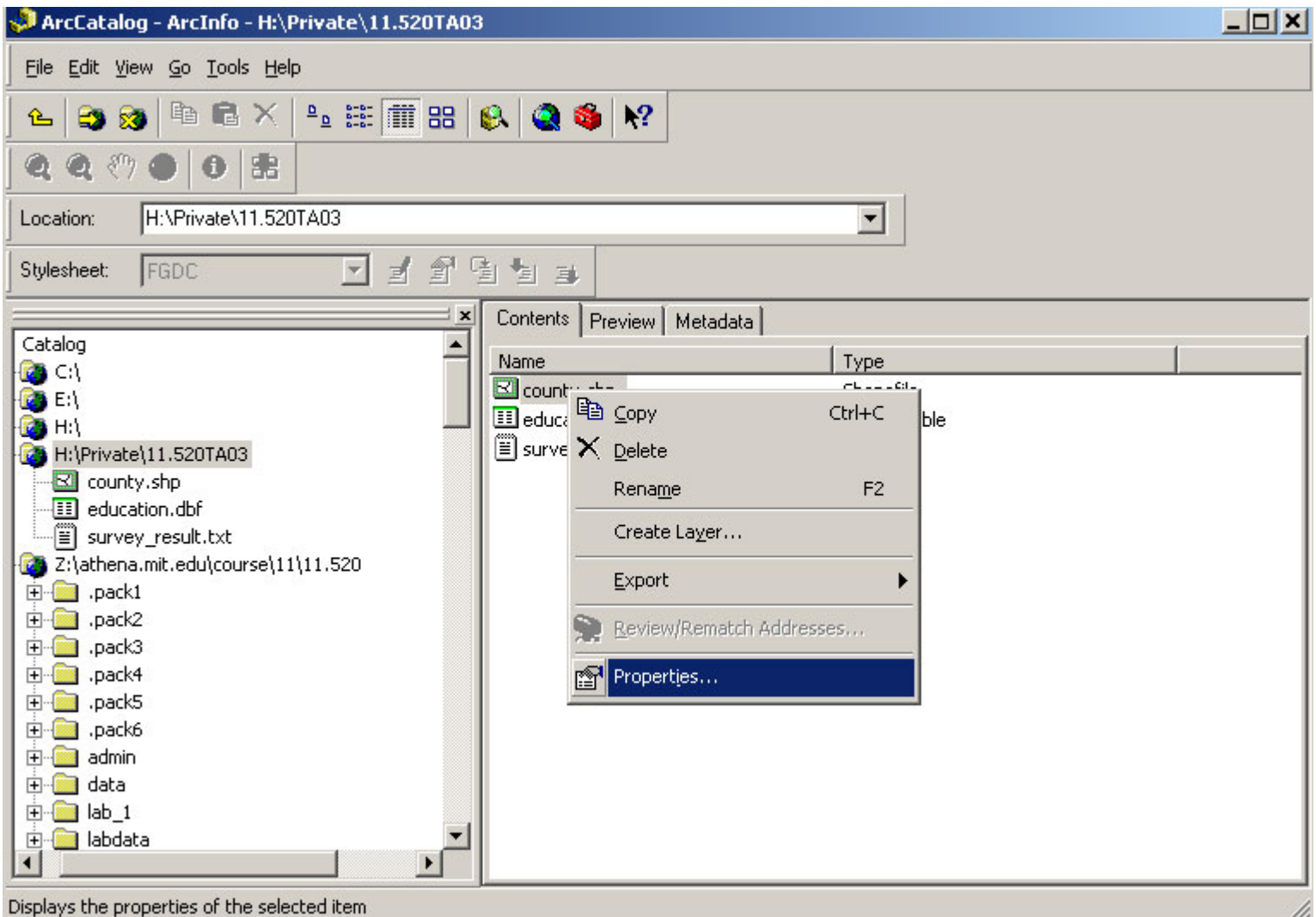


Fig. 9. Open Shapefile Properties Window

2) Click the "Geometry" from the Data Type list. Field properties information will show up in the bottom of the **Shapefile Properties** window. Click the "... " button in the right side of the "Spatial Reference" property. It will bring up **Spatial Reference Properties** window

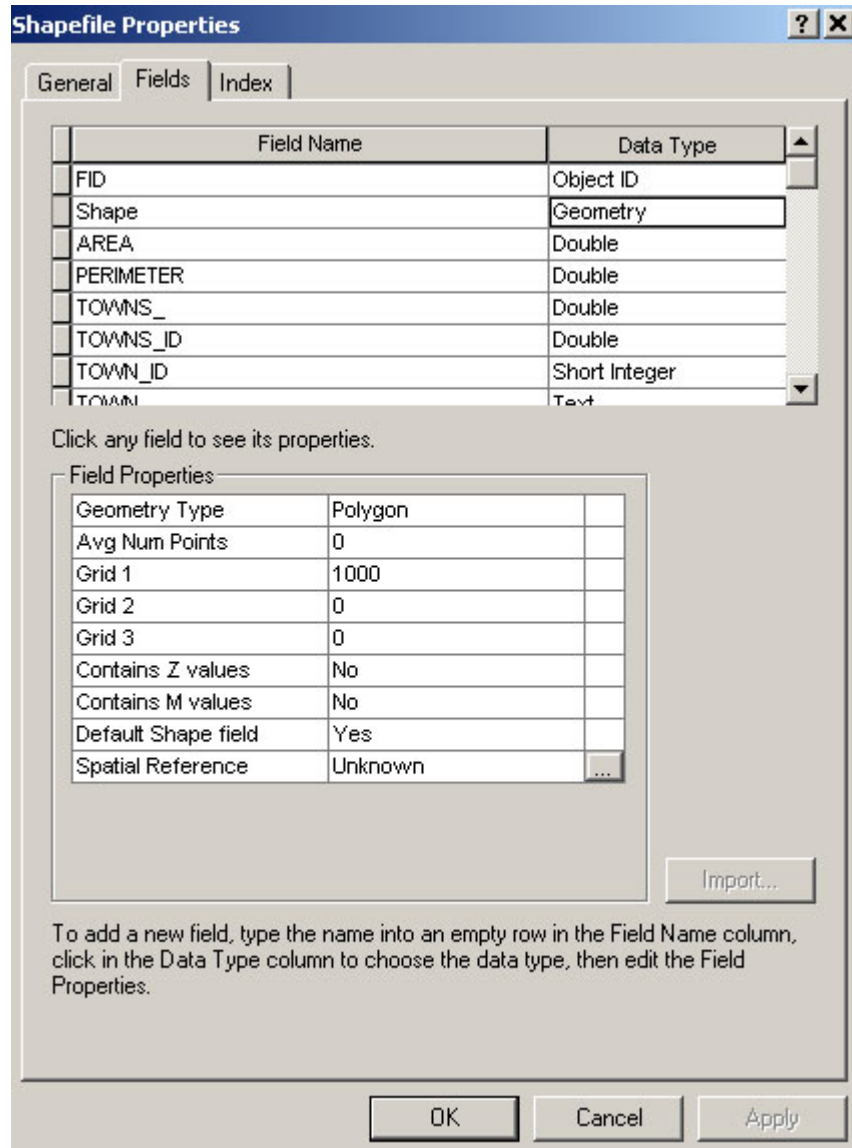


Fig. 10. shapefile Properties Window

3) Click the **Select** button from the Spatial Reference Properties Window.

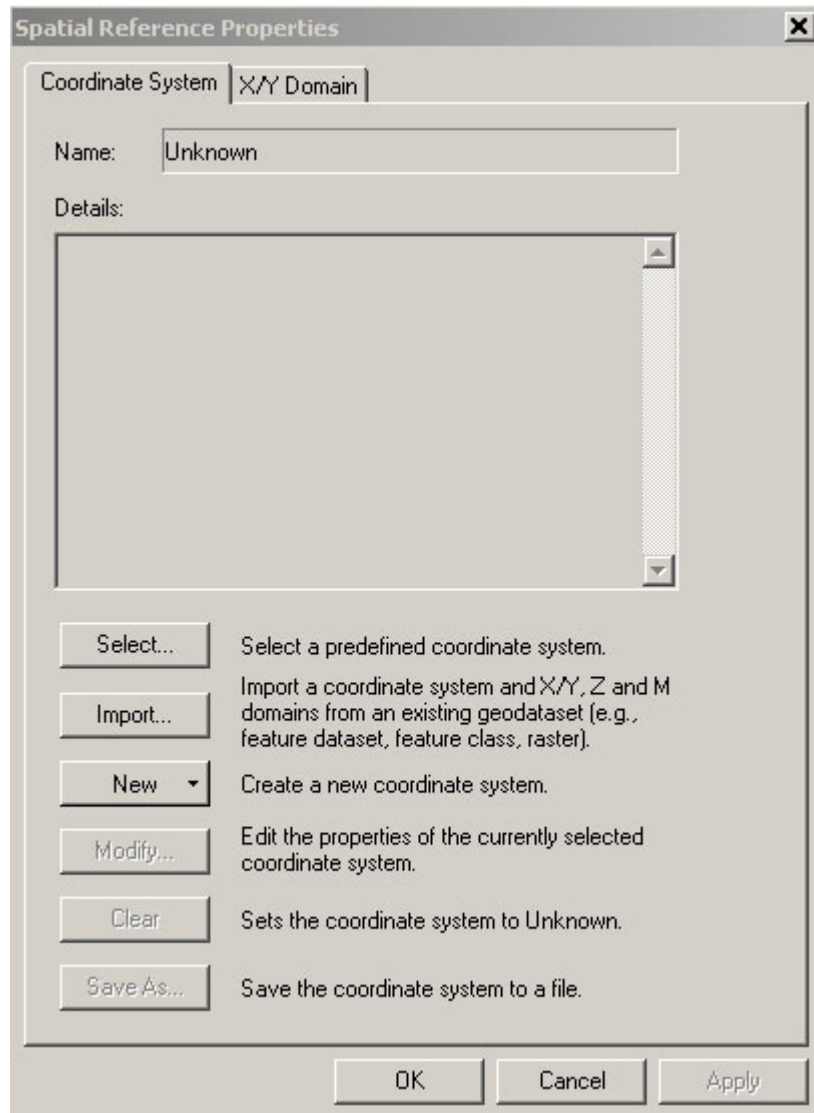


Fig. 11. Spatial Reference Properties Window

4) Select appropriate spatial reference system from the list. In this case **NAD1983 Stateplane Massachusetts Mainland FIPS 2001.prj** is the right one. Select the spatial reference system and click **ADD**.

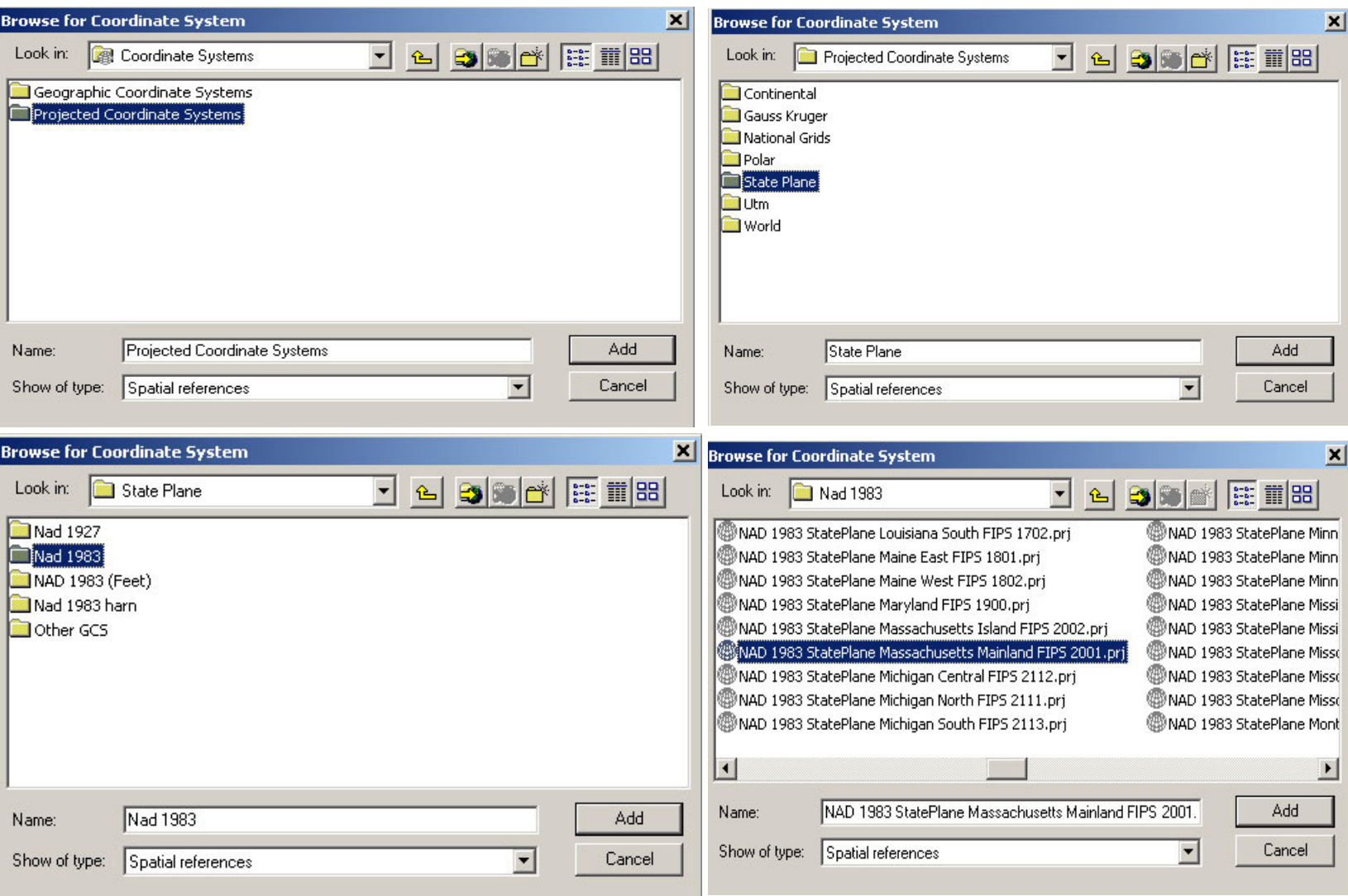


Fig. 12. Select spatial reference system

5) Now the spatial reference for the **county** layer has been set. Click **OK**. Now go to the ArcMap and add the **county** layer to the **legend map** dataframe. This time no warning message will show up.

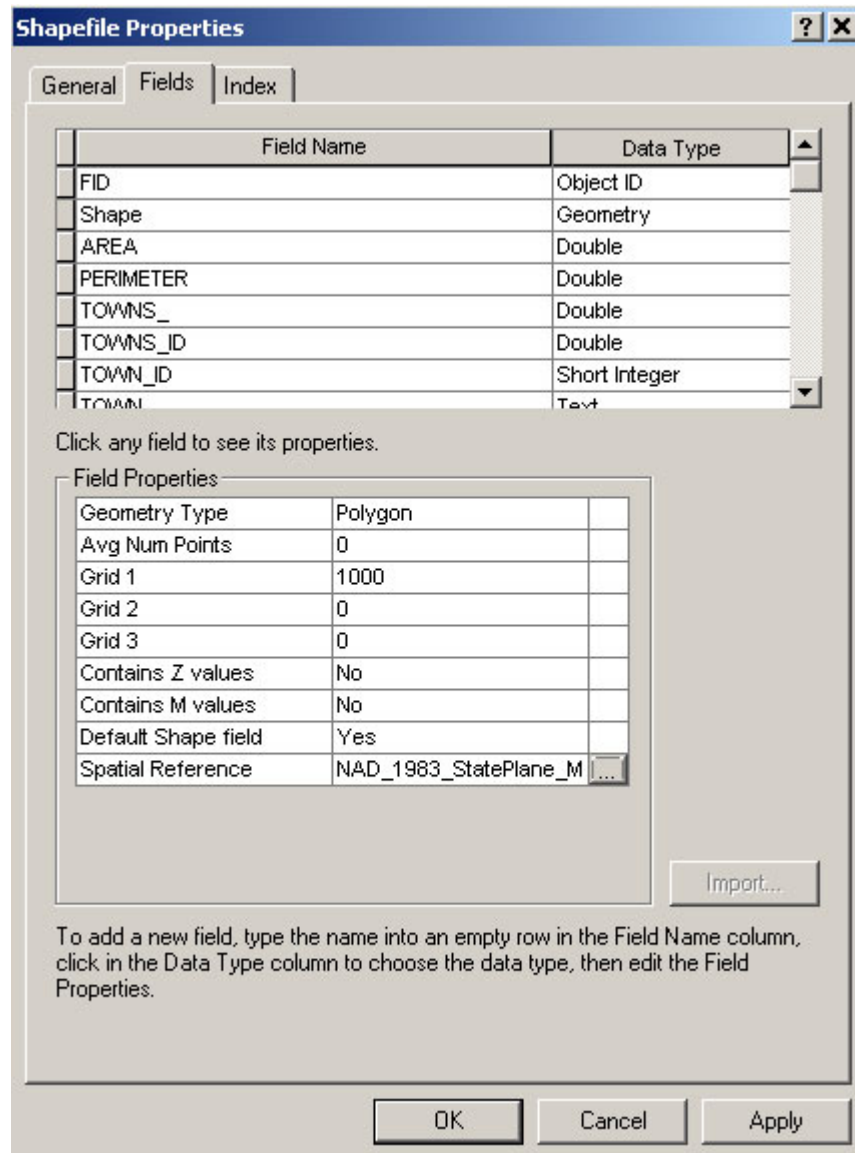


Fig. 13. Set Spatial Reference System

Knowing how to examine and reset the projection associated with data frames and layers is often handy when integrating data layers that come from different agencies and sources.

IV. Lab Assignment

Today's lab assignment has 6 questions. Please turn in the answer sheet to Xiongjiu Liao. The lab is due in **Lab 4**.
