

11.520: A Workshop on Geographic Information Systems

11.188: Urban Planning and Social Science Laboratory

Lab Exercise 7: Raster Spatial Analysis

Distributed: Lab 7

Due: Lab 8

Overview

The purpose of this lab exercise is to introduce spatial analysis methods using raster models of geospatial phenomena. Thus far, we have represented spatial phenomena as discrete features modeled in the GIS as points, lines, or polygons--i.e., so-called 'vector' models of geospatial features. Sometimes it is useful to think of spatial phenomena as 'fields' such as temperature, wind velocity, or elevation. The spatial variation of these 'fields' can be modeled in various ways including contour lines and raster grid cells. In this lab exercise, we shall focus on **raster** models and examine ArcGIS's '**Spatial Analyst**' extension.

We shall use raster models to create a housing value 'surface' for Cambridge. A housing value 'surface' for Cambridge would show the high- and low-value neighborhoods much like an elevation map shows height. To create the 'surface' we will explore ArcGIS's tools for converting vector data sets into raster data sets--in particular, we will 'rasterize' the 1989 housing sales data for Cambridge and the 1990 Census data for Cambridge block groups.

The block group census data and the sales data contain relevant information about housing values, but the block group data may be too coarse and the sales data may be too sparse. One way to generate a smoother housing value surface is to interpolate the housing value at any particular location based on some combination of values observed for proximate housing sales or block groups. To experiment with such methods, we will use a so-called 'raster' data model and some of the ArcGIS Spatial Analyst's capabilities.

The computation needed to do such interpolations involve lots of proximity-dependent calculations that are much easier using a so-called 'raster' data model instead of the vector model that we have been using. Thus far, we have represented spatial features--such as Cambridge block group polygons--by the sequence of boundary points that need to be connected to enclose the border of each spatial object--for example, the contiguous collection of city blocks that make up each Census block group. A raster model would overlay a grid (of fixed cell size) over all of Cambridge and then assign a numeric value (such as the block group median housing value) to each grid cell depending upon, say, which block group contained the center of the grid cell. Depending upon the grid cell size that is chosen, such a raster model can be convenient but coarse-grained with jagged boundaries, or fine-grained but overwhelming in the number of cells that must be encoded.

In this exercise, we only have time for a few of the many types of spatial analyses that are possible using rasterize data sets. Remember that our immediate goal is to use the `cambbgrp` and `sales89` data to generate a housing-value 'surface' for the city of Cambridge. We'll do this by 'rasterizing' the block group and sales data and then taking advantage of the regular grid structure in the raster model so that we can easily do the computations that let us smooth out and interpolate the housing values.

I. Setting Up Your Work Environment

1. Launch the ArcGIS and add five data layers listed below.

- `M:\data\cam_lu99.shp` Cambridge land use in 1999 per MassGIS
- `M:\data\cambbgrp_point.shp` Census 1990 block group' centroids for Cambridge
- `M:\data\cambbgrp.shp` Census 1990 block group polygons for Cambridge
- `M:\data\cambtigr (coverage)` U.S. Census 1990 TIGER file for Cambridge
- `M:\data\sales89` Cambridge Housing Sales Data
- `M:\data\camborder` Cambridge polygon

2. Set **Display unit = meter**. In this exercise you will use "**Meter**" instead of using "**Mile**"

II. Spatial Analyst Setup

ArcGIS's raster manipulation tools are bundled with its **Spatial Analyst** extension. It's a big bundle so lets open ArcGIS's help system first to find out more about the tools. Open the ArcGIS help page by clicking **Help > ArcGIS Desktop help** from the menu bar. Click the **index** tab and type "**Spatial analyst**". During the exercise, you'll find these online help pages helpful in clarifying the choices and reasoning behind a number of the steps that we will explore. Be sure, at some point, to take a look at the **Overview** section.

The Spatial Analyst module is an extension, so it must be loaded into ArcGIS separately. To load the Spatial Analyst extension:

- Click the **Tools** menu
- Click '**Extensions**' and check '**Spatial Analyst**'
- Click '**Close**'

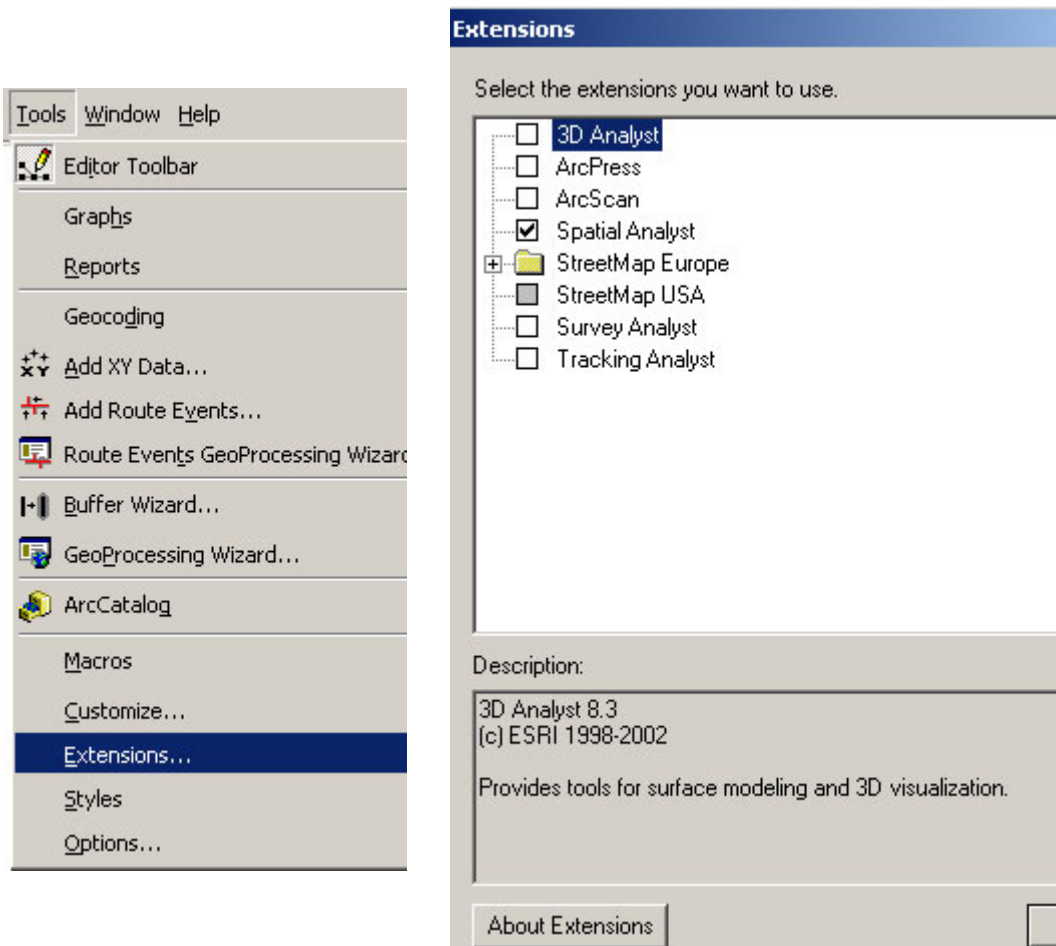


Fig. 1. Add Extension

Although you just activated the "**Spatial Analyst**" extension, you have to add the "**Spatial Analyst** tool bar on the menu manually to use the extension (quite

inconvenient!!!). To add "*Spatial Analyst*" tool bar, go to **View > Tool bars** from the menu bar and click "*Spatial Analyst*"

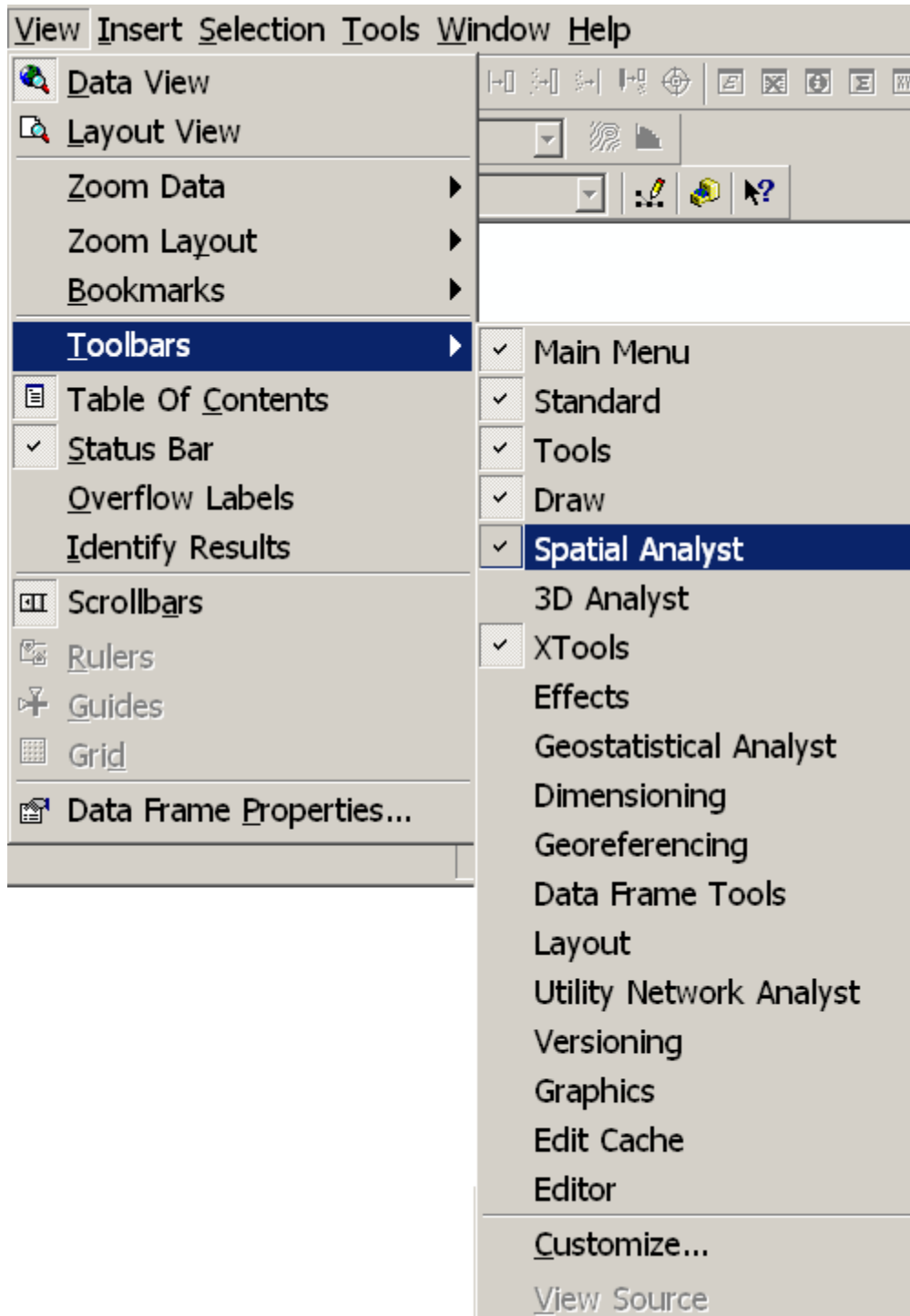


Fig. 2. Add Spatial Analyst toolbar

Once the Spatial Analyst tool bar loaded, a new main-menu heading called *Spatial Analysis* will be available whenever you launch the ArcGIS.

Setting Analysis Properties:

Before building and using raster data sets, we should set the grid cell sizes, the spatial extent of our grid, and the 'no data' regions that we wish to 'mask' off. Let's begin by specifying a grid cell size of 100 meters and an analysis extent covering all of Cambridge. To do this, click '**Spatial Analyst > Option**'. When the "**Options**" window pops up.

- In the **General** tab, set your working directory, select *None* for the Analysis mask (We will set the mask later), and select the *first option* for the Analysis Coordinate System.
- In the **Extent** tab, select *Same as Layer "camborder polygon"* for the Analyst extent.
- In the **Cell Size** tab, select *As Specified Below* then specify
 - Cell size--100
 - Number of rows--57
 - Number of columns--79

(Number of rows and Number of columns will be automatically computed.)

Now that we've set the analysis properties, we are ready to cut up Cambridge into 100-meter raster grid cells. Convert the **camborder** to a grid layer using these steps and parameter settings:

- Select **Spatial Analyst > Convert > Feature to Raster. Features to Raster** window will show up.
 - Choose *camborder polygon* for the **Input features**.
 - Choose **COUNTY** for the **Field**.
(We just want a single value entered into every grid cell at this point. Using the County field will do this since it is the same across Cambridge)
 - Output cell size should be **100**
 - Set the saving location (**your working directory**) and the name of the grid file (*cambordergd*) and click **OK**.

If successful, the **CAMBORDERGD** layer will be added to the data frame window. Turn it on and notice that the shading covers all the grid cells whose center point falls inside of the spatial extent of the **camborder** layer. The cell value associated with the grid cells is 25017--the FIPS code number for the county. Since we did not join feature attributes to the grid, there is only one row in the attribute table for **CAMBORDERGR** - attribute tables for raster layers contain one row for each unique grid cell value - hence, there is only one row in this case.

At this point, we don't need the old **camborder** coverage any longer. We used it to set the spatial extent for our grid work, but that setting is retained. To reduce clutter, you can delete the **camborder** layer.

III. Interpolating Housing Values Using SALES89

This part of the lab will demonstrate some techniques for filling in missing values in your data using interpolation methods. In this case, we will explore different ways to estimate housing values for Cambridge. Keep in mind that there is no perfect way to determine the value of a property.

A city assessor's database of all properties in the city would generally be considered a good estimate of housing values because the data set is complete and maintained by an agency which has strong motivation to keep it accurate. This database does have drawbacks, though. It is updated sporadically, people lobby for the lowest assessment possible for their property, and its values often lag behind market values by many years or even decades.

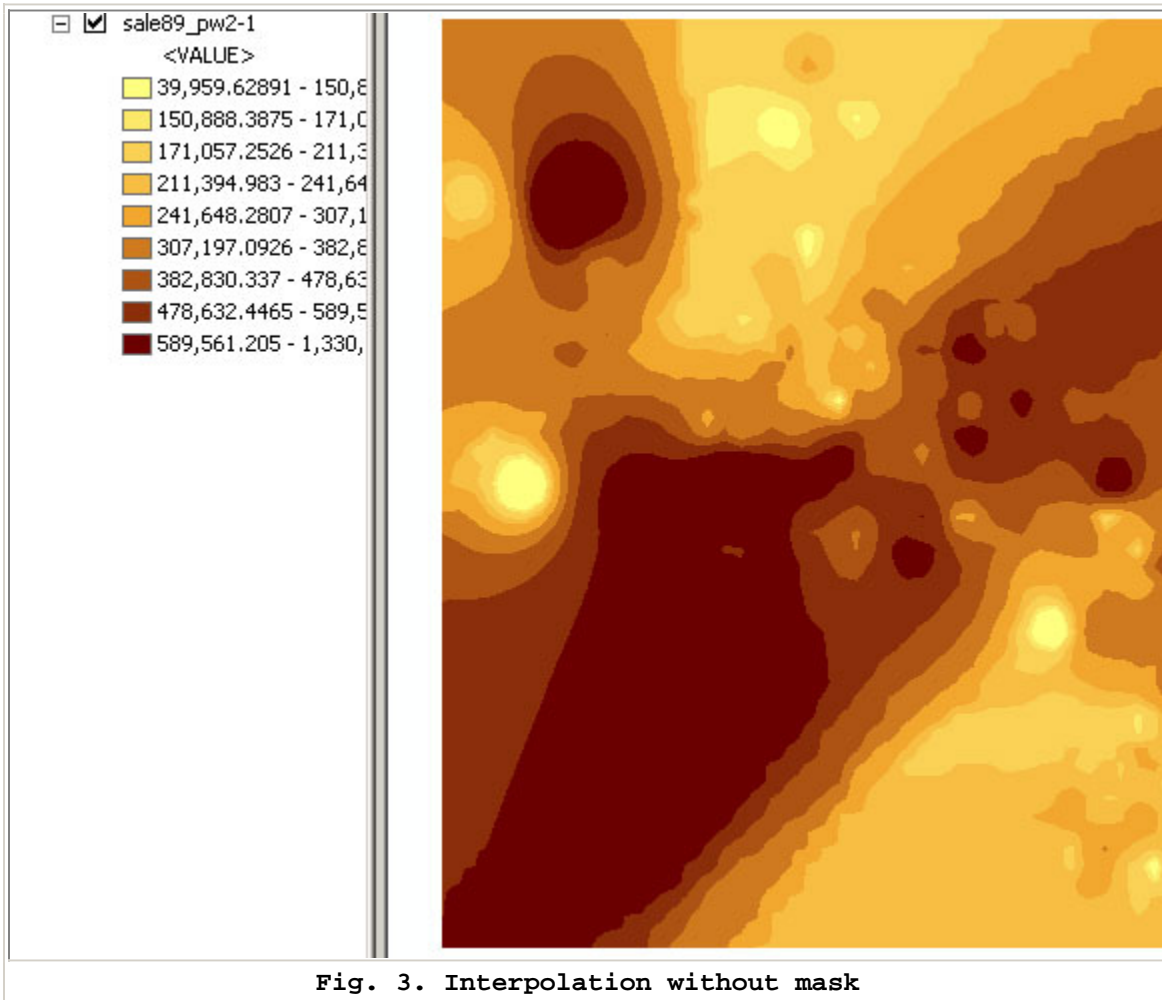
Recent sales are another way to get at the question. On the one hand, their numbers are believable because the price should reflect an informed negotiation between a buyer and a seller that results in the 'market value' of the property being revealed (if you are a believer in the economic market-clearing model). However, the accuracy of such data sets are susceptible to short-lived boom or bust trends, not all sales are 'arms length' sales that reflect market value and, since individual houses (and lots) might be bigger or smaller than those typical of their neighborhood, individual sale prices may or may not be representative of housing prices in their neighborhood.

Finally, the census presents us with yet another estimate of housing value--the median housing values aggregated to the block group level. This data set is vulnerable to criticism from many angles. The numbers are self-reported and only a sample of the population is asked to report. The benefit of census data is that they are cheaply available and they cover the entire country.

We will use `sales89` and `cambbgrp` to explore some of these ideas. Let's begin with `sales89`.

- Be sure your data frame contains at least these layers: `sales89`, `cambbgrp`, and `cambordergd`.
- Select '*Spatial Analyst > Interpolate to Raster > Inverse Distance Weighted*'. Specify options when **Inverse Distance Weighted** window shows up
 - Input points: "`sales89 point`".
 - Z value field: "`REALPRICE`"
 - Power: "`2`"
 - Search radius type: "**Variable**"
 - Number of points: "`12`"
 - Maximum distance: leave it blank
 - Uncheck the "**Use barrier polylines**:"
 - Output cell size: "`100`"
 - Output raster: "`[your working directory]/sale89_pw2-1`"

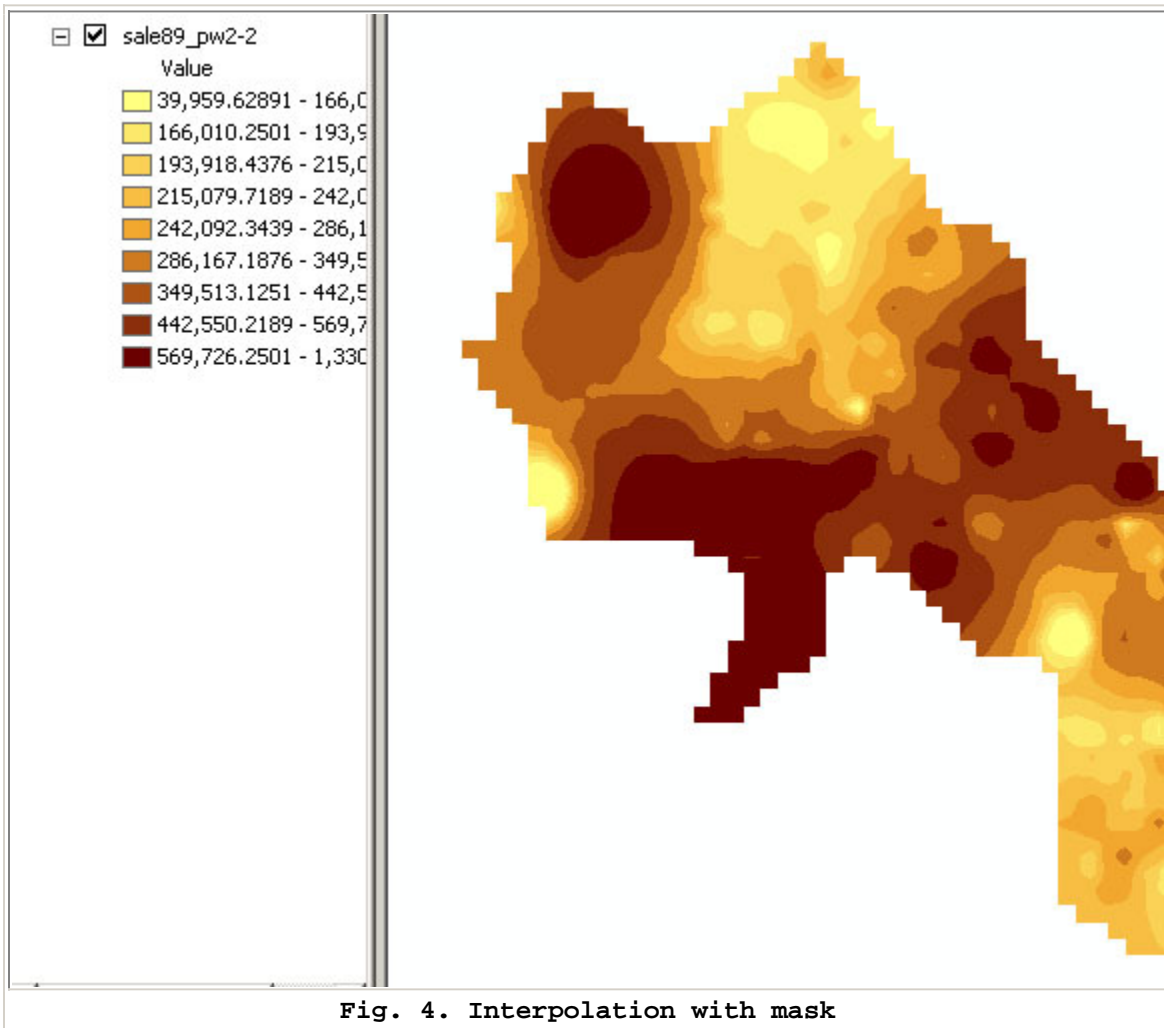
- Click OK
- The grid layer that is created fills the entire bounding box for Cambridge and looks something like this:
(When using **Quantile, 9 classes**)



The interpolated surface is shown thematically by shading each cell dark or light depending upon whether that cell is estimated to have a lower housing value (darker shades) or higher housing value (lighter shades). Based on the parameters we set, the cell value is an inverse-distance weighted average of the 12 closest sales. Since the power factor was set to the default (2), the weights are proportional to the square of the distance. This interpolation heuristic seems reasonable, but the surface extends far beyond the Cambridge borders (all the way to the rectangular bounding box that covers Cambridge). We can prevent the interpolation from computing values outside of the Cambridge boundary by 'masking' off those cells that fall outside of Cambridge. Do this by adding a mask to the Analysis Properties:

- Reopen the **Spatial Analysis > Options** dialog box and set the **Analysis Mask** to be **CAMBORDERGD** (the grid that we computed earlier from the **camborder** coverage).

With this analysis mask set, interpolate the **Realprice** values in **sales89** once again and save it as **sales89_pw2-2**. The **sales89_pw2-2** should look like this:
(When using **Quantile, 9 classes**)



All the values inside Cambridge are the same as before, but the cells outside Cambridge are 'masked off'.

To get some idea of how the interpolation method will affect the result, redo the interpolation (using the same mask) with the power set to 1 instead of 2. Label this surface '**Sales89-pw1**'. Use the **identify** tool to explore the differences between the values for the point data set, **sales89**, and the two raster grids that you interpolated. You will notice that the grid cell has slightly different values than the realprice in **sales89**, even if there is only one sale falling within a grid cell. This is because the interpolation process looks at the city as a continuous value surface with the sale points being sample data that

gives an insight into the local housing value. The estimate assigned to any particular grid cell is a weighted average (with distant sales counting less) of the 12 closest sales (including any within the grid cell). In principle, this might be a better estimate of typical values for that cell than an estimate based only on the few sales that might have occurred within the cell. *On your lab assignment sheet, write down the original and interpolated values for the grid cell in the upper left that contains the most expensive Realprice value in the original sales89 data set.* Do you understand why the interpolated value using the power=1 model is considerably lower than the interpolated value using the power=2 model? There was only one sale in this cell and it is the most expensive 1989 sale in Cambridge. Averaging it with its 11 closest neighbors (all costing less) will yield a smaller number. Weighting cases by the square of the inverse-distance-from-cell (power=2) gives less weight to the neighbors and more to the expensive local sale compared with the case where the inverse distance weights are not squared (power=1).

Finally, create a third interpolated surface, this time with the interpolation based on all sales **within 1000 meters** and **power=2** (rather than the 12 closest neighbors). To do this, you have to set the **Search radius type: Fixed** and **Distance: 1000** in the **Inverse Distance Weighted dialog box**. Call this layer '**sales89-1000m**' and use the identify tool to find the interpolated value for the upper-left cell with the highest-priced sale. (Confirm that the display units are set to meters in **View > Data Frame Properties** before interpolating the surface. The distance units of the view determine what units are used for the distance that you enter in the dialog box.) *What is this interpolated value and why is this estimate even higher than the power=1 estimate?*

Note: None of these interpolation methods is 'correct'. Each is plausible based on a heuristic algorithm that estimates the housing value at any particular point to be one or another function of 'nearby' sales prices. The general method of interpolating unobserved values based on location is called 'kriging' and the field of spatial statistics studies how best to do the interpolation (depending upon explicit underlying models of spatial variation). See, for example, *S+ Spatial Stats: User's Manual for Windows and Unix*, by Kaluzny, Bega, Cardoso, and Shelly, MathSoft, Inc., 1998 (ISBN: 0-387-98226-4) for further discussion of kriging techniques using the 'Spatial Statistics' add-on of a high-powered statistical package called S+ that is available on Athena.

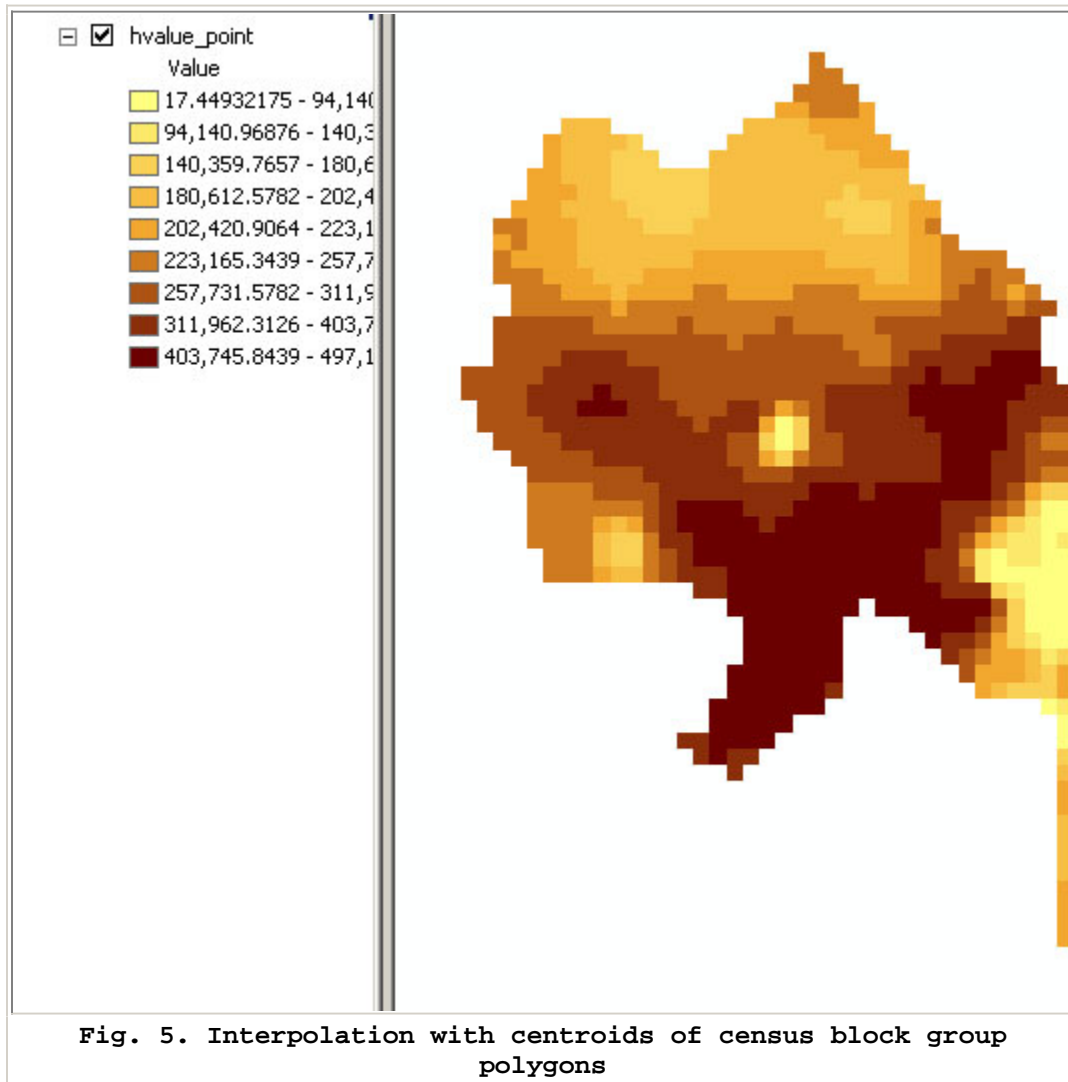
IV. Interpolating Housing Values Using CAMBBGRP

Another strategy for interpolating a housing value surface would be to use the median housing value field, **med_hvalue**, for the census data available in **cambbgrp**. There are several ways in which we could use the block group data to interpolate a housing value surface. One approach would be exactly analogous to the **sales89** method. We could assume that the block group median was an appropriate value for some point in the '**center**' of each block group. Then we could interpolate the surface as we did above if we assume that there was one house sale, priced at the median for the block group, at each block group's '**center**' point. A second approach would be to treat each block group median as an

average value that was appropriate across the entire block group. We could then rasterize the block groups into grid cells and smooth the cell estimates by adjusting them up or down based on an average housing value of neighboring cells.

Let's begin with the first approach.

- Select *Spatial Analyst > Interpolate to Raster* and choose 'Inverse Distance Weighted'.
- Select *cambgrp_point* as your input layer and *med_hvalue* as your Z Value Field. Take the defaults for method, neighbors, and power.
- Name this layer *hvalue_point*.
- Click *O.K* and you should get a shaded surface like this:
(When using **Quantile, 9 classes**)



Next, let's use the second approach (using the polygon data) to interpolate the housing value surface from the census block group data.

- If you haven't already done so, add the `cambbgrp.shp` to your data frame.
- Select ***Spatial Analyst > Convert > Feature to Raster***. **Features to Raster** window will show up.
 - Choose ***cambbgrp*** for the **Input features**.
 - Choose ***MED_HVALUE*** for the **Field**.
 - Output cell size should be **100**
 - Set the saving location (**your working directory**) and the name of the grid file (***cambbgrpgd***) and click **OK**.

As you can see from the below images, except for the jagged edges, the newly created grid layer looks just like a vector-based thematic map of median housing value.

(When using **Quantile, 9 classes**)



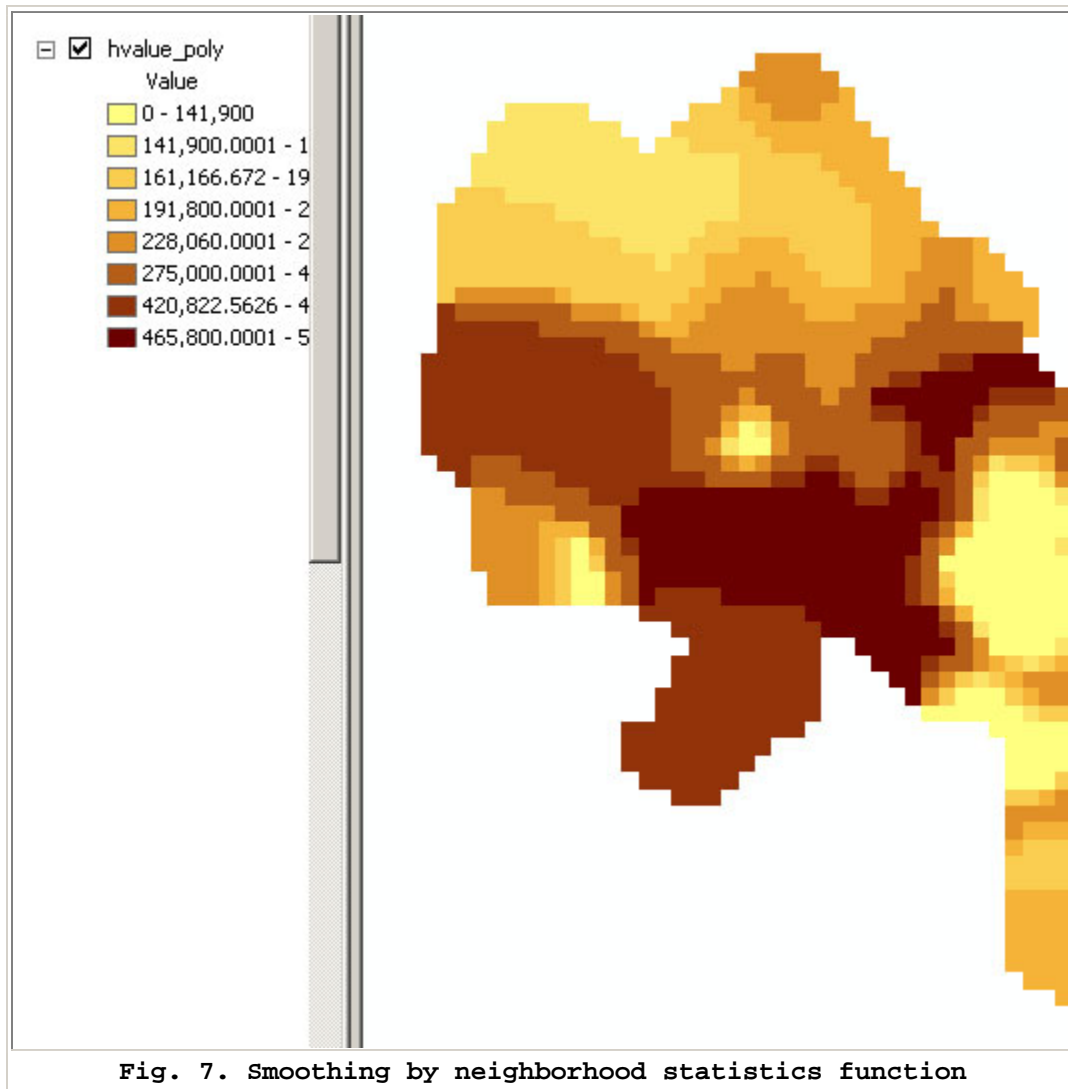
Examine its attribute table. It has 63 unique values (except 0) --one for each unique value of **med_hvalue** in the original `cambbgrp` coverage. The attribute table for grid layers contains one row for each unique value (as long as the cell value is an integer and not a floating point number!) and a **count** column is included to indicate how many cells had that value. Grid layer layers such as `hvalue_points` have floating point values for their cells and, hence, no attribute

table is available. (You could reclassify the cells into integer value ranges if you wished to generate a histogram or chart the data.)

Finally, let's smooth this new grid layer using the **Spatial Analyst > Neighborhood Statistics** option. Let's recalculate each cell value to be the average of all the neighboring cells - in this case we'll use the 9 cells (a 3x3 matrix) in and around each cell. To do this, choose the following settings: (they are the defaults)

- Input data: **CAMBBGRPGD**
- Field: **Value**
- Statistic type : **Mean**
- Neighborhood:**Rectangle**
- Width: **3**
- Height: **3**
- Units: **cell**
- Output cell size: **100**
- Output raster: **[your working space]/hvalue_poly**

Click '**OK**' and the hvalue_poly layer will be added on your data frame. Change the classify method to "Quantile". You should get something like this:
(When using **Quantile, 8 classes**)

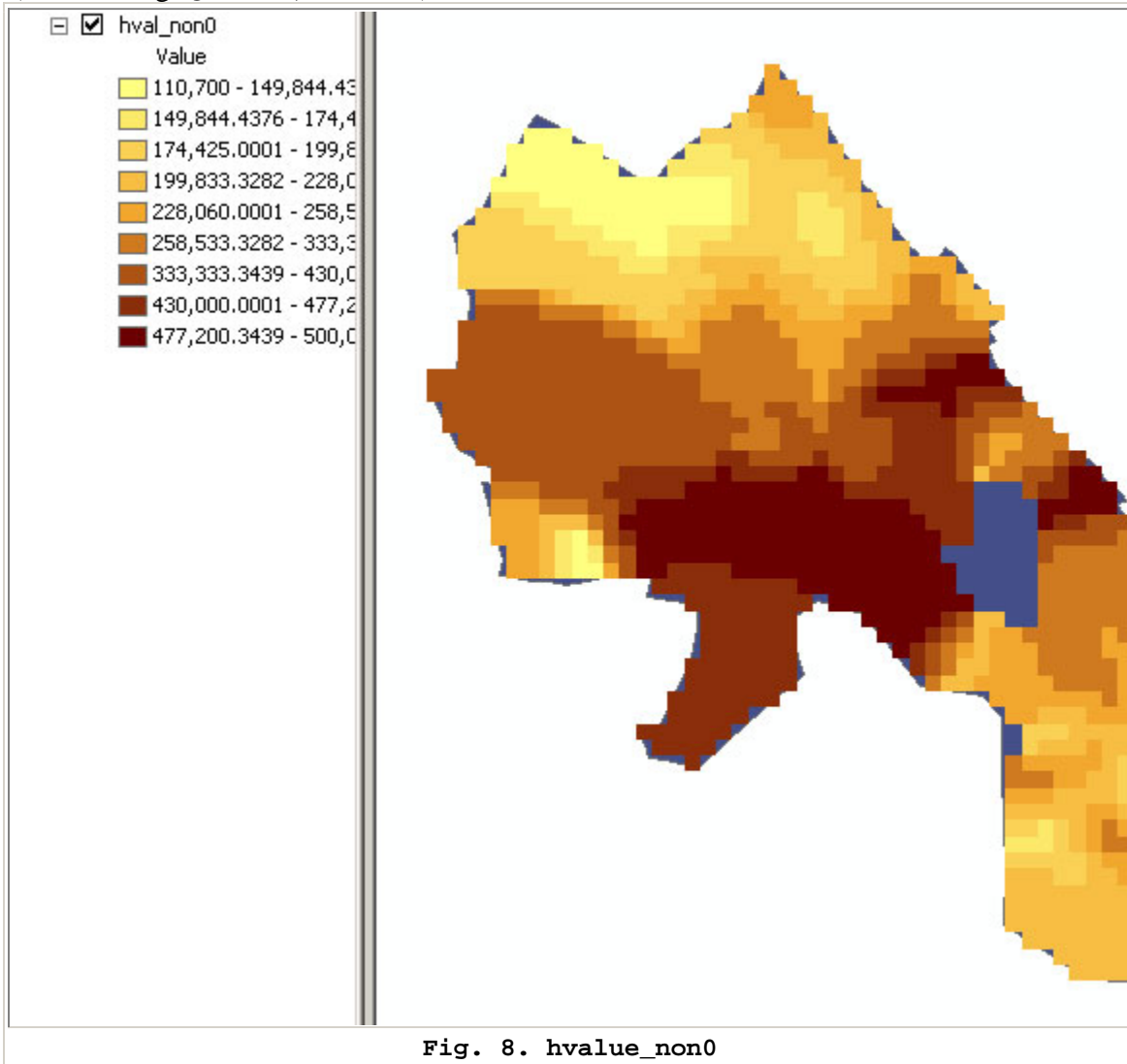


Note that selecting rows in the attribute table (for hvalue_poly) will highlight the corresponding cells on the map. *Find the cell containing the location of the highest price sales⁸⁹ home in the northwest part of Cambridge. What is the interpolated value of that cell using the two methods based on med_hvalue?*

Many other variations on these interpolations are possible. For example, we know that **med_hvalue** is zero for several block groups--presumably those around Harvard Square and MIT where campus and commercial/industrial activities results in no households residing in the block group. Perhaps we should exclude these cells from our interpolations--not only to keep the 'zero' value cells from being displayed, but also to keep them from being included in the neighborhood statistics averages. Copy and paste the **cambbgrp.shp** layer and use the query tools in the **Layer Properties > Definition Query** tab to **exclude** all block groups with **med_hvalue = 0** (which means include all block groups with med_hvalue > 0). Now, recompute the polygon-based interpolation (including the neighborhood averaging) and call this grid layer '**hvalue_non0**'. Select the

same color scheme as before. In the data window, turn off all layers except the original `cambordergd` layer (displayed in a non-grayscale color like blue) and the new `hvalue_non0` layer that you just computed. The resulting view window should look something like this:

(When using **Quantile, 9 classes**)



Notice the no-data `cambordergd` cells sticking out from under the new surface and notice that the interpolated values don't fall off close to the no-data cells as rapidly as they did before (e.g., near Harvard Square). You'll also notice that the low-value categories begin above \$100,000 rather than at 0 the way they did before. This surface is about as good an interpolation as we are going to get using the block group data.

Comment briefly on some of the characteristics of this interpolated surface of `med_hvalue` compares with the ones derived from the `sales89` data. Are the hot-spots more concentrated or diffuse? Does one or another approach lead to a broader range of spatial variability?

V. Combining Grid Layers Using the Map Calculator

Finally, let us consider combining the interpolated housing value surfaces computed using the sales89 and med_hvalue methods. ArcGIS provides a '**Raster calculator**' option that allows you to create a new grid layer based on a user-specified combination of the values of two or more grid cell layers. Let's compute the simple arithmetic average of the **sale89_pw2-2** grid layer and the **med_hvalue_non0** layer. Select *Spatial Analyst > Raster Calculator* and enter this formula:

$([\text{hvalue_non0}] + [\text{sale89_pw2-2}]) / 2$

and click Evaluate.

The result is a new grid which is the average of the two estimates and looks something like this:

(When using **Quantile, 9 classes**)

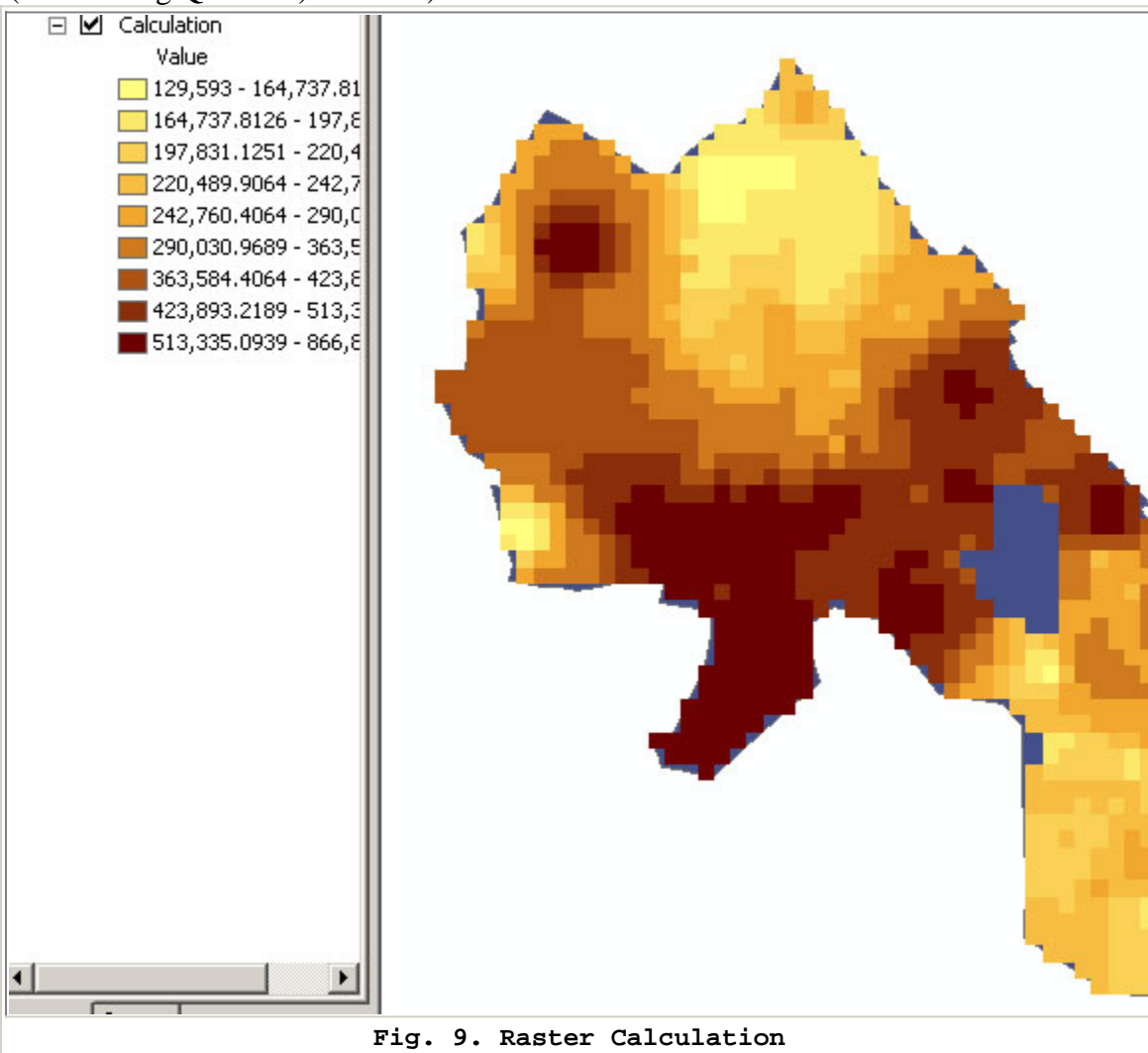


Fig. 9. Raster Calculation

The map calculator is a powerful and flexible tool. For example, if you felt the sales data was more important than the census data, you could assign it a higher weight with a formula such as:

$([\text{Med_hvalue}] * 0.7 + ([\text{Sales_Price}] * 1.3)) / 2$

The possibilities are endless--and many of them won't be too meaningful! Think about the reasons why one or another interpolation method might be misleading, inaccurate, or particularly appropriate. For example, you might want to compare the mean and standard deviation of the interpolated cell values for each method and make some normalization adjustments before combining the two estimates using a simple average. ***For the lab assignment, however, all you need do at this point is determine the final interpolated value (using the first map-calculator formula) for the cell containing the highest price sales⁸⁹ house (in the Northwest corner of Cambridge). Write this value on the assignment sheet.***

We have only scratched the surface of all the raster-based interpolation and analysis tools that are available. If you have extra time, review the help files regarding the Spatial Analyst extension and work on those parts of the homework assignment that ask you to compute a population density surface for youths.
