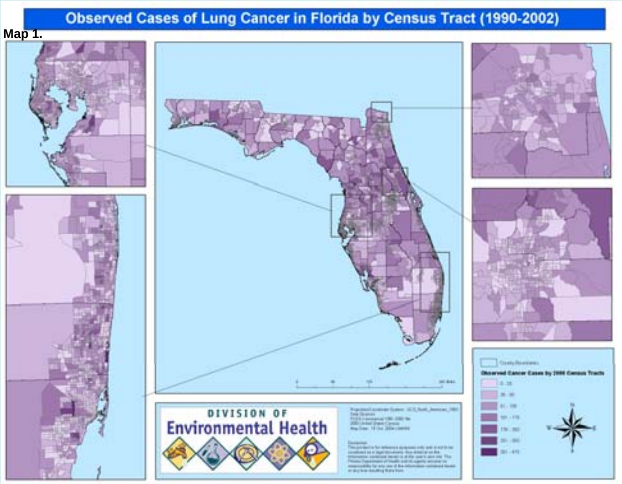
Multiple Visualizations of Lung and Bronchus Cancer in Florida by Census Tract

Laurel Harduar-Morano MPH, Florida Department of Health

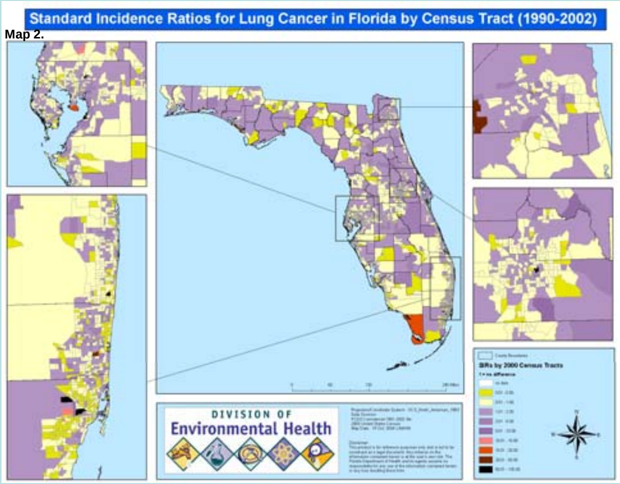
The purpose of this poster is to provide a few examples of chronic disease, i.e. lung cancer, data visualization. It is important for public health officials to be able to generate hypotheses of areas with potential problems where further investigation is required. Mapping of aggregated data can provide a good starting point.

Initially, a map of Florida was created that only examined the observed cases per census tract. This method is a highly inaccurate way of visualizing disease data – the map does not take into account varying population size and age distribution between census tracts. For instance five cases of lung cancer in a population of 100,000 is not unlikely while 5 cases in a population of 100 may suggest a problem. Due to this reason it is impossible to compare results from one census tract to another census tract using only observed cases.

Map 1.



Map 2.



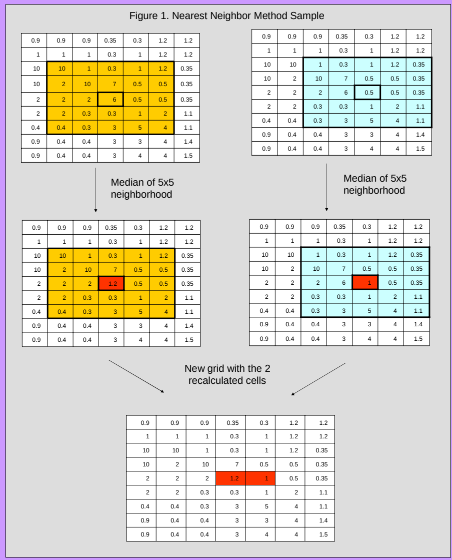
A map utilizing age-adjusted Standard Incidence Ratios (SIRs) allows for the comparability of each census tract. SIRs are a method of indirect standardization which, in this case, compares the number of cases observed in each census tract to a number of expected cases in the same tract, which were calculated by multiplying age-specific rates of lung cancer from the overall Florida population and the total population for the same census tract.

SIR values less than 1 indicate that less cases of lung cancer than expected where found in the census tract

SIR values greater than one indicate that more cases of lung cancer than expected where found in the census tract

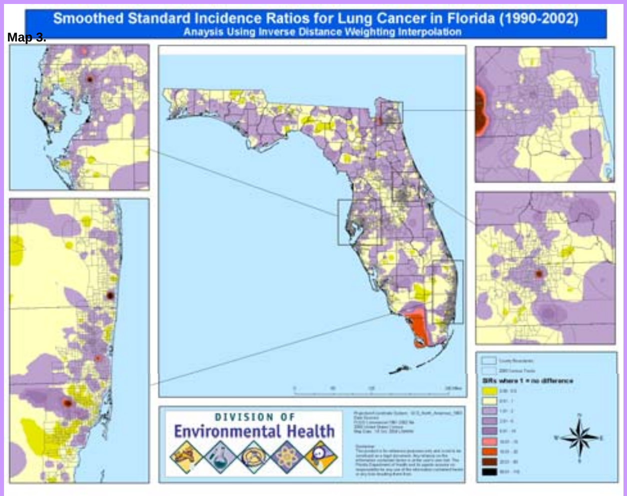
SIR values equal to one indicate that there was no difference between observed cases of lung cancer and expected cases of lung cancer.

Maps of age-adjusted SIRs, while a big step up from mapping observed cases, have their own problems, they may obscure the spatial pattern of disease distribution, especially since the rates are based on census tracts with differing populations. Rates which are based on small populations or small numbers of observed cases are more likely to be artificially elevated. For instance, in Florida where the age-specific rate of lung cancer is 3%; a census tract population with approximately 8,120 individuals and 48 observed cases of lung cancer would have an SIR of 0.42, 58% lower than expected. Another census tract with 900 individuals, the same age distribution and number of observed cases would have an SIR of 3.81, 281% above the expected.

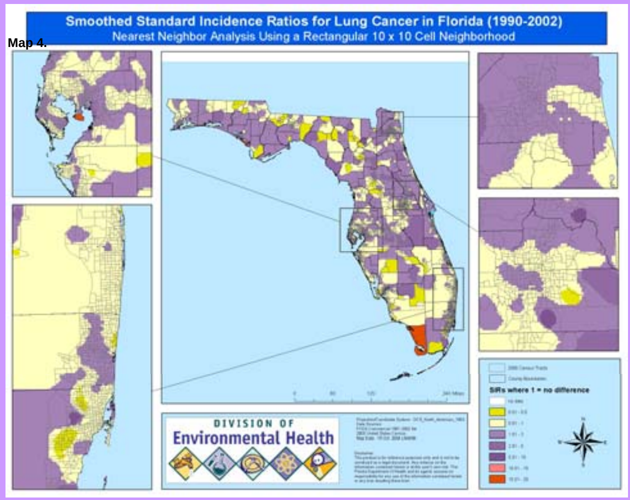


Spatial smoothing, particularly nearest neighbor analysis is one method utilized to create a more accurate picture of the overall spatial pattern of disease risk. ArcMap's Spatial Analyst extension was used to convert features, polygons of 2000 census tracts, into a raster. A layer file which contained the Florida state boundary was used as a mask to avoid trying to convert areas with no data into raster format. The size of each cell, 2,300 m^2, was determined to be approximately half the size of the smallest census tract. In this way a grid was created with each cell containing an SIR value relative to its census tract. The raster file was then used in the nearest neighbor analysis. The median SIR value for an input cell centered within a rectangular *y x y* cell neighborhood (or a circular neighborhood with a radius of y cells) was calculated for each cell in the raster, as demonstrated to the left.

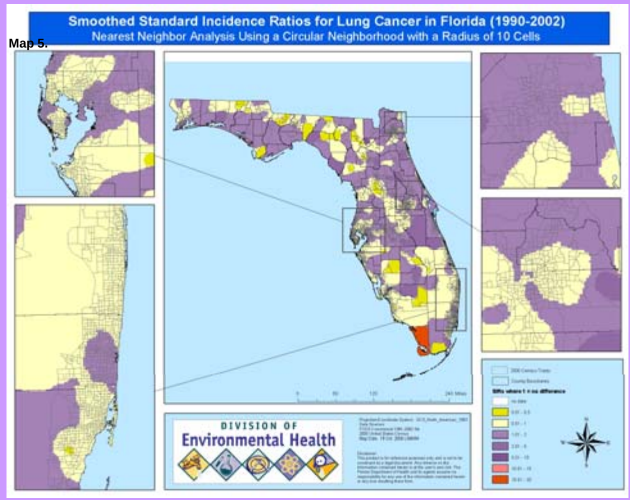
Map 3.



Map 4.

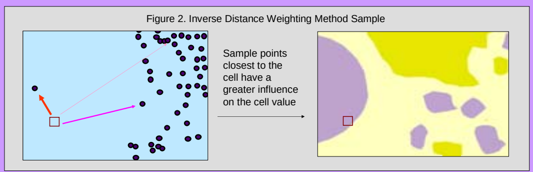


Map 5.



A second smoothing technique weights sample points relative to the estimated cell. This method is Inverse Distance Weighting. IDW is an interpolation method which creates a continuous surface from sample points. Utilizing the Spatial Analyst extension it produces a raster file as its output.

The layer file which contained the Florida state boundary was again used as a mask to avoid trying to convert areas with no data into raster format. The size of each cell was the same as before, 2,300 m^2 approximately half the size of the smallest census tract. The centroids of each census tract polygon were the inputted sample points. Each centroid had the same value as their census tract polygon. The IDW method estimates each cell value by averaging the values of the sample points, census tract centroids, within the cells neighborhood. In this case the cell neighborhood consisted of 12 sample points. The census tract centroids closest to the cell, within the neighborhood, had a greater weight in the calculation and the most influence on the cell value while the centroids farther away from the cell had a smaller weight and less of an influence on the cell value. This is shown in the figure to the left.



All three of the above maps (maps 3 - 5) show varying degrees of smoothing. For each map the level of smoothing will change depending on the size of the neighborhood and the type of calculation utilized.

For instance, when comparing the nearest neighbor maps; there is a great deal of color variation in the map that used the 10 x 10 rectangular neighborhood (map 4), while the circular neighborhood with a radius of 10 cells (map 5) reduced a lot of the color variation This is due to the fact that the circular neighborhood coves a larger square meter range than the rectangular neighborhood. To determine which of the two maps to use depends on whether the objective requires a more global or localized visualization of the spatial pattern of chronic disease distribution.

The second model, IDW, also used a neighborhood approach. The most important difference in this model compared to the first model, nearest neighbor, is that the second model used a weighted average instead of the median to estimate the cell values. The first and second model both address the issue of census tracts as unrealistic boundaries. They both provide a good visualization of the spatial pattern of risk. However, by using an average instead of the median the resulting map, created by the IDW model (map 3), is sensitive to the extreme SIR observations and the skewed distributions. The weighted median provides a method which is better adapt at reducing the image *noise*, the arbitrary effect.

The purpose of the final map is to allow the audience to visualize the spatial pattern of lung cancer distribution within the state of Florida. By smoothing the SIRs we were trying to reduce the noise in the image created by artificially elevated SIRs and the high variance in the census tract populations. In addition, since individuals do not live/work according to census tract boundaries, there needed to be a way to estimate the overlapping areas between the boundaries. All this was accomplished by borrowing information from parts of neighboring census tracts (or multiple centroids). Overall spatial smoothing is highly variable and subjective in the end as with all maps it depends on the type of information that is to be presented. However, all of the maps provide valuable information for hypothesis generating. However, for the purposes of mapping public health data, a nearest neighbor analysis of age-adjusted SIRs with a weighted median would provided the most tempered and accurate results.