Project #1: Spatial Analysis

1.1 Problem Statement

High nitrate concentrations in drinking water are a health hazard. Recently a possible cancer risk in adults from nitrate (and nitrite) has emerged, but the magnitude of the risk is unknown. Imagine that you work for the Wisconsin Department of Natural Resources. The agency has collected data on cancers by cataloging the location of all cancer occurrences over a ten-year period. In addition, they assembled a database of nitrate levels in a group of test wells throughout the county.

Your job is to search for a relationship between nitrate level and cancer. Bear in mind that the well locations have nothing to do with cancer occurrence. Rather, the well data come from required water testing when new wells are dug or when existing wells are repaired by licensed contractors. In other words, the nitrate data points are more-or-less randomly distributed. The cancer data and nitrate data can be found in the associated shapefiles.

Write a program that uses inverse distance weighting ($1/d^k$) to produce a raster map of nitrate levels. There is no theory to say what the distance exponent k should be, thus your program needs to be able to make a map for any value $k > 1$. Maps produced by your program will be used to see if high levels of nitrate are related to cancer locations. In order to demonstrate that your results are robust, you must try a reasonable range of k values.

Test for the existence of clustering in cancers for the county as a whole. Explain why that is an obvious first test for any possible role of nitrate in cancer incidence. Write a program implementing kernel density estimation to compute the density of cancers in occurrences per square kilometer. You may either use an adaptive kernel method, or use a fixed kernel whose bandwidth is appropriate for the data. Use a defendable measure of “high density” to identify clusters of cancers. Compare this to the pattern of nitrate levels. Prepare a report of not more than five pages explaining your analysis methods and the findings in terms that an educated but non-technical person can understand.

1.2 Conceptualization

1.2.1 Problem analysis

The core of the problem is to compare the spatial distribution of nitrate and cancer. Typically, there are many methods to examine the relationship between two spatial phenomena, like spatial overlay or spatial regression. They all require that we have some kind of measurement for both of them at the same location. Right now, both of the data layers we have – cancer incidences and nitrate from test wells – are discrete point data and the well locations have nothing to do with cancer occurrence, which means they are not necessarily at the same locations. So we need to first generate layers that have measurement for both of them over the entire region, like creating a continuous surface over the region or aggregating them to some spatial units that cover the entire region.

1.2.2 Measurement of Nitrate

We all know that the nitrate level in the ground water change continuously over the landscape, so the distribution of nitrate can be represented by a continuous surface over the region. But how can we generate the surface from a set of well observations?

In G579, we learned an important spatial analysis method – interpolation, which can be used to predict the values of a variable at every point in space using the known data values of sample points. It looks just like what we need to generate the surface of nitrate. But note that interpolation is based on the assumption that spatially distributed objects are spatially correlated; in other words, things that are close together tend to have similar characteristics. It should not be applied to objects that do not satisfy the assumption. We can use interpolation here because nitrate level changes continuously over the region. The well locations can be used as the sample points.

1.2.3 Measurement of Cancer

Unlike nitrate, cancer incidences are discontinuous. It occurs to individual persons and does not change continuously over the region. So it would be difficult to represent it in the same way as nitrate level.

However, we can aggregate the variable, for example, count cancer incidences to some form of spatial units. Note that a high number of incidences does not mean a high rate of cancer. Cancer occurs among people, which means in most cases, the more people there are, the more cancer incidences you would see. The best way to measure the prevalence might be to weight it by the population. Check the “population-at-risk” of G579 for more information of this part.

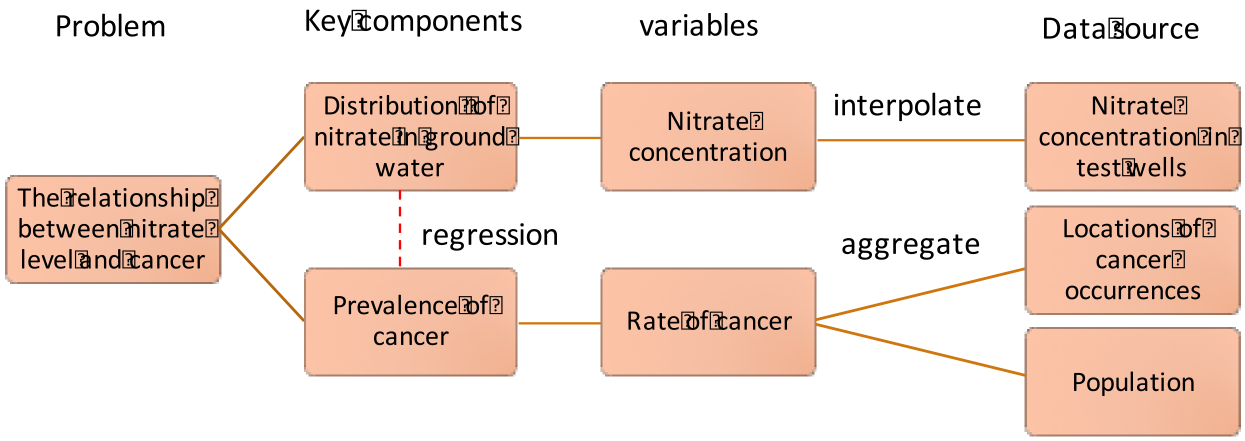
Another thing to consider is the selection of spatial unit. There are different levels of spatial units, e.g. counties, census tracts, blocks. Since the study region only covers one state, the county level might be too general to show the variation of the rate of cancer over the region, or to show the relationship between the two phenomena. So we probably want to use census tracts or blocks.

1.2.4 Find the relationship

Now that we have the continuous field of nitrate and the rate of cancers aggregated to some form of spatial units, we need to compare them on the same level. So we need to aggregate nitrate concentration to the same spatial unit of cancer rates. For example, we can calculate the average nitrate concentration in each block. After that, we can perform spatial regression to find the relationship between them.

1.3 Conceptualization diagram

From the analysis above, we can come up with the following conceptualization diagram. This represents the general analysis of the problem. There are still many issues to resolve in order to implement it, and we will examine some of them in the next section.


Problem
Key components
variables
Data source
The relationship between nitrate level and cancer
Distribution of nitrate in ground water
Nitrate concentration
interpolate
Nitrate concentration in test wells
regression
Prevalence of cancer
Rate of cancer
aggregate
Locations of cancer occurrences
Population

1.4 Issues

1.4.1 Spatial Interpolation

In G579, we talked about many spatial interpolation techniques, like IDW, kriging and spline. IDW is a relatively simple method, and is also easy to implement. Kriging is more restrictive and requires large amount of sample data. It also requires that the stationarity assumption holds. So kriging might not be the proper solution. Spline is best when we want a smooth surface. It could be an option here, but it is might be too complicated to implement for this project.

In this project, we will use IDW for the interpolation (but feel free to experiment with other methods). One of the key issues of IDW is the choice of distance decay coefficient. The distance decay coefficient q determines how fast weight will decrease as distance increases. The choice of q is important for a proper interpolation. There is no theory to say what the distance decay coefficient should be; thus your program needs to be able to make a map for any value $q > 1$. Maps produced by your program will be used to see if high levels of nitrate are related to cancer locations. To demonstrate that your results are robust, try a reasonable range of k values and use what you leaned form G579 to justify your selection.

1.4.2 Aggregation

Cancer incidences will be aggregated to the block group level (typically 1-2K people per block group). Rather than providing a crude rate of cancer occurrences per block group, we will provide a more meaningful definition that takes into account population. Therefore, cancer rate is defined as total cancer occurrences divided by the total population of the block group.

1.4.3 Regression

There are many regression methods. In this project, we are going to perform linear regression in order to simplify the implementation. Linear regression analyses attempt to demonstrate the degree to which one or more variables potentially promote positive or negative change in another variable.

1.5 Implementation Guideline

Step 1: Gather datasets on cancers and nitrate levels and geocode the datasets.

In this case, the students need to decide how to examine the relationship between cancer and nitrate levels, how to quantify the spatial distribution of cancers and nitrate levels.

You are provided with two shapefiles for this project: Cancer.shp and Well.shp. Cancer.shp contains the rate of cancer aggregated at census tract level. Well.shp contains the nitrate level at sample wells.

Step 2: Develop a workflow and determine your solution stack. The final deliverable must be a program that allows the user to adjust parameters like q to produce a static or dynamic map to visualize the results. Below are a few ideas to get you started:

> 1. Python Program: This project can be accomplished fully in ArcMap (and it is suggested you try the workflow manually in ArcMap first). Thus, writing a script in python using ArcPy should be fairly easy. If you go this route, you must provide some sort of GUI in place of the website where the user can change parameters like q, view, and download the map. There are a number of python GUI libraries out there.
>
> 2. Web Framework: You could use a web framework like flask or node.js to initiate open source spatial libraries or even a python script to produce a web map.
>
> 3. Traditional Web: With this option you could place the data in a database and use spatial queries to create results. Another option could be AGOL coupled with the Arc Javascript api.
>
> There are many other routes to go. It is advised you choose a workflow you are most comfortable with, but also challenges your skillset.

Step 3: Research and setup the development environment for the solutions to the issues above (relating spatial distribution of nitrate levels and cancer rates through programming) and laying out the steps for implement (the steps to program the analysis of relationship).

> *Example: You may need to setup Tomcat, configure it for Eclipse Development, install GDAL for basic spatial analysis, and develop client-side (JSP, Leaflect, Dojo, etc.).*

Step 4: Tie the solutions together to complete the initial implementation (such as a browser-based interface to the implementation

Step 5: Test and enhance the implementation, complete project report.