Task: GIS exploration of 1854 London cholera data

The purpose of this exercise is to undertake an exploratory analysis of John Snow's classic 1854 London cholera dataset using GIS and to reflect on the research questions raised, that are covered in greater detail elsewhere in this teaching material. Here, it is more important to think through the ways in which two point patterns may be related, rather than the achievement of a specific statistical description of the relationships.

Throughout the exercise, there are four questions to answer that appear in boxes. These questions form the basis for your assignment. The instructions assume that you are working with Spatial Analyst in ArcGIS and we assume that you are reasonably familiar with some of the Spatial Analyst commands already. If you feel you would like more explanation on any of the commands, please ask through the message boards.

Data:

The data packaged for this exercise consist of two shape files: **Cholera_Deaths** and **Pumps**, taken from Robin Wilson's web site at http://blog.rtwilson.com/john-snows-cholera-data-in-more-formats/, where you will also find relevant background information. There are also scanned georeferenced images of the modern day street pattern in Soho, London. The original data themselves were derived from the dataset created by Tobler (1994) at http://www.ncgia.ucsb.edu/pubs/snow/snow.html, although the data are no longer downloadable from that site.

"The data consists of the … the location of 578 deaths from cholera, and the position of 13 water pumps (wells). Each coordinate point in the file 'deaths' specifies the address of a person who died from cholera…. The dates of the deaths are not recorded."

Tobler (1994)

NB John Snow's maps were, of course, hand-drawn and had no explicit coordinate or projection system. The data we will use for this exercise have been georeferenced to the Ordnance Survey National Grid.

Exercise:

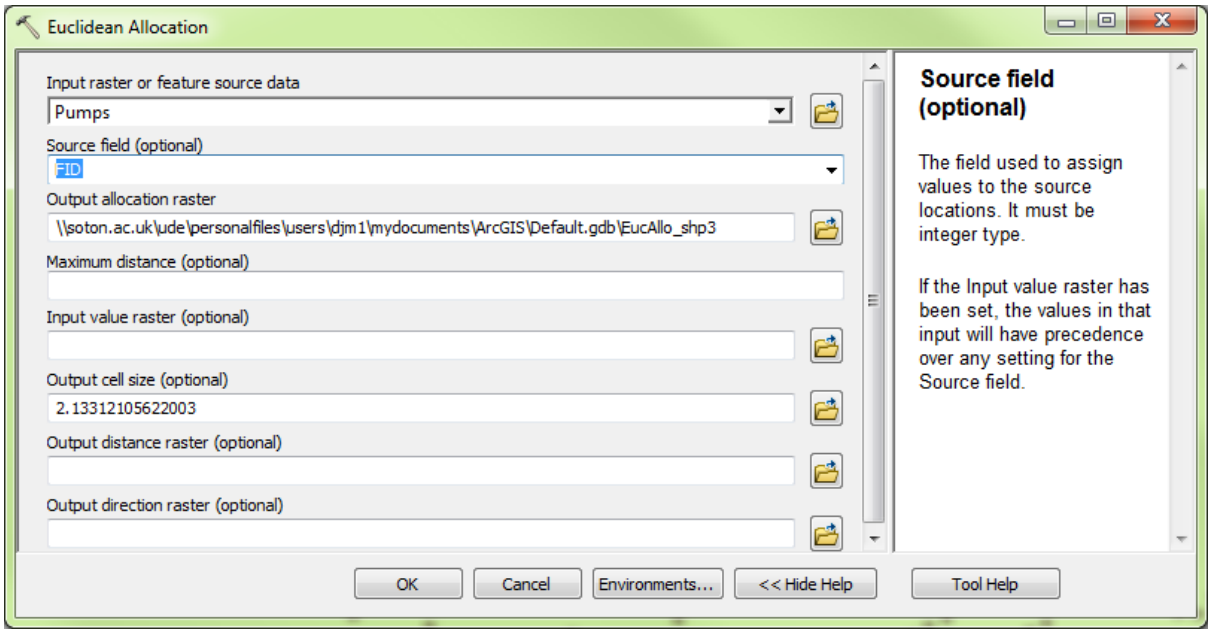
Display each of the data layers and perform an initial visual analysis of the data provided. To display the deaths, you may wish to right-click on the **Cholera_Deaths** layer, select properties, then choose the symbology tab and select quantities under show. If you select 'graduated symbols' and then under value, you can select the **count** field – this contains a count of the number of deaths recorded at a given address

. You may find the resultant display more useful for the deaths.

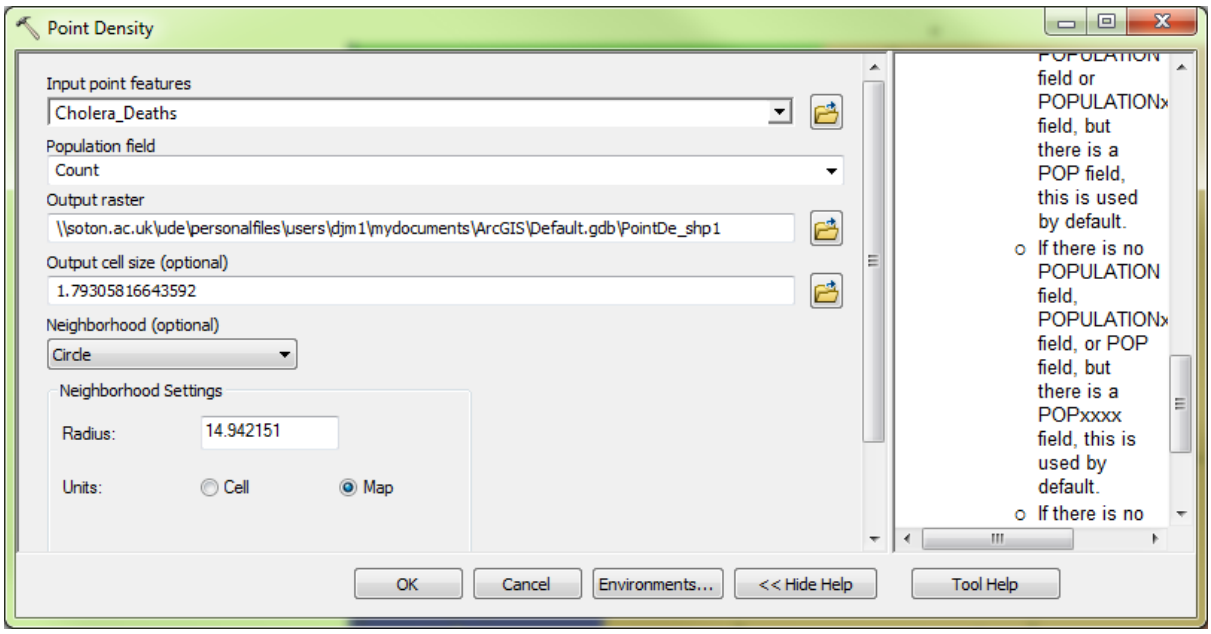
1. Disregarding any prior knowledge or background reading, does the pattern observable in the datasets suggest the presence of any particular associations between pump locations and cholera cases?

We can begin our analysis from the perspective of the pumps (i.e. the potential causal locations for the disease), which was the approach followed by John Snow:

Use the Help documentation to read about the *Euclidean Allocation* function in Spatial Analyst (available in the ArcToolBox, under *Spatial Analyst tools / distance*). When running Euclidean Allocation based on **Pumps** ensure that **FID** is selected as the source field. The resulting image shows locations that are closer to each pump than to any other – this type of polygon is frequently referred to as a Thiessen or Voroni tessellation and provides a rough idea of the area served by each pump. Now display the **Cholera_Deaths** layer over your new allocation raster.



Use the Help documentation to review the operation of the *Point density* function and try running it on the **Cholera_Deaths** dataset, again comparing your results with the original point distributions. So that the number of deaths at each address is incorporated into the estimate of deaths per area, be sure to specify count as the population field (e.g. this will mean that where there were 2 deaths at an address, both will be counted when calculating the local density of deaths). The default search radius is unlikely to be the most useful, so explore the effect of using different search radii.



Zonal Statistics can be used to summarise the values of one raster layer within the zones of another dataset. Choose one of your point density maps and then use the Zonal Statistics tool to find the mean density of **Cholera_Deaths** for each of the **Pumps** Euclidean allocation zones that you created earlier.

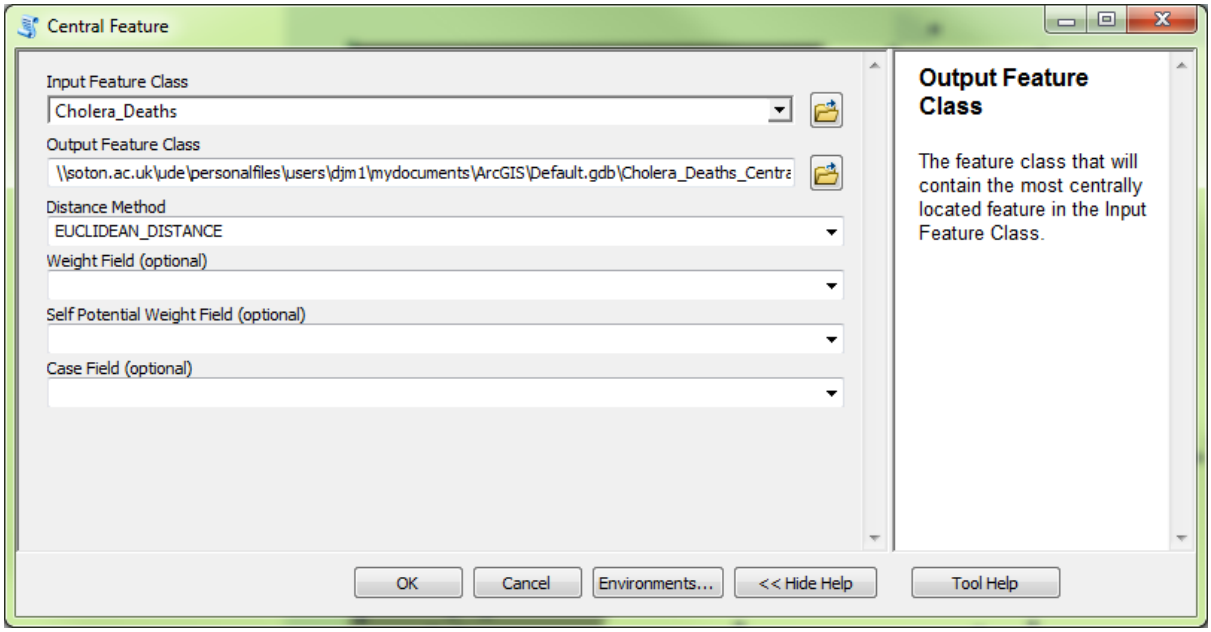
2. How would you interpret the results from the Zonal Statistics command? Do the results of Zonal Statistics confirm the association apparent from simply eye-balling the data?

An alternative approach to the spatial analysis of this dataset is to begin with the pattern of deaths. Snow has often been credited with this analysis, although it is not a true reflection of his work as he already suspected the water supply to be implicated in the cholera outbreak.

This time, we will begin our analysis by trying to summarise the pattern of deaths.

Within the ArcToolBox, go to *Spatial Statistics Tools, Measuring Geographic Distributions*, and take a look at the help for the *Mean Center* command. Try running this command, using the **Cholera_Deaths** as the *input feature class*.

Now go back to *Spatial Statistics Tools, Measuring Geographic Distributions*, and take a look at the help for the *Central Feature* command. Try running this command, again using the **Cholera_Deaths** as the *input feature class*.



3. How would you interpret the results of the *Central Feature* and *Mean Center* tools? Do they support the findings from our earlier analysis?

Contemporary spatial analysis offers many more approaches for quantifying the concepts illustrated here, but the underlying issues remain the same and are just as relevant to GIS analysis of a modern disease dataset as to this historical disease outbreak.

4. Consider the representation of the pumps, street network and death locations. Identify as many ways as possible in which the characteristics of the data available here may have unintended effects on the results of any spatial analysis applied. Do you think there are any limitations in the analysis that we have undertaken?