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 consist of two shape files: cholera_deaths and pumps, taken from Robin Wilson's web site (see html for details). There are also scanned georeferenced images of the modern day street pattern in Soho, London. The original data themselves were derived from the dataset downloadable from Tobler (1994) Snow's cholera map at http://www.ncgia.ucsb.edu/pubs/snow/snow.html

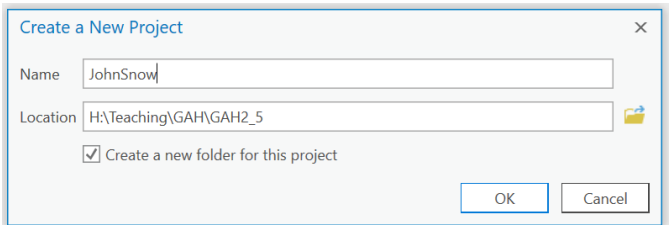
The data consists of the … the location of 578 deaths from cholera, and the position of 13 water pumps (wells). Each coordinate point in the file "deaths" specifies the address of a person who died from cholera…. The dates of the deaths are not recorded.

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. See Robin Wilson's blog for more: http://blog.rtwilson.com/john-snows-cholera-data-in-more-formats/.

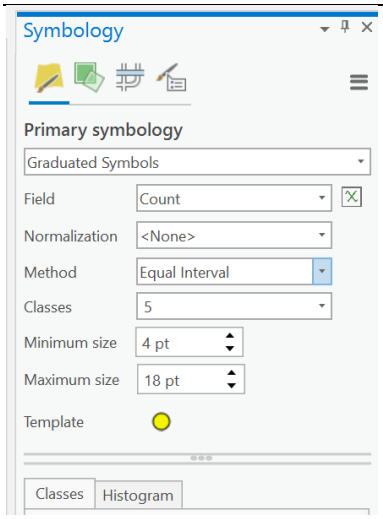
Exercise:

If you are new to ArcPro, start up the software and choose a new map template. Choose a suitable name and location for your project.



Having set up your new project, to load data, you first need to set up a connection to one or more folders. To do this, right-click on 'folders' in the catalog window on the right, then '*add folder connection'* to specify the folder containing the map layers for this exercise. You can display layers from within your chosen folder by dragging and dropping them into the central map canvas window.

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 *symbology*, and select *graduated symbols* under *primary symbology.* 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. Note that graduated symbols vary point symbol size in discrete 'bands' derived from death counts; proportional symbols vary each point's area in proportion to the count of deaths for that point. You may wish to experiment with proportional symbols in addition to graduated symbols:

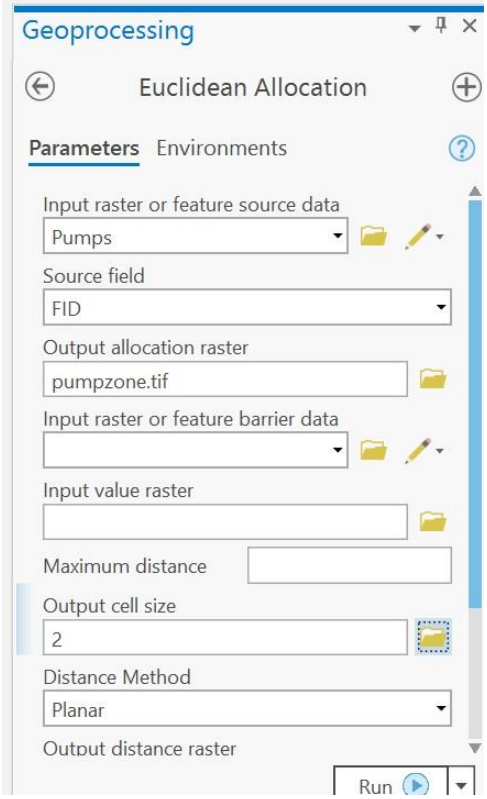


1. Disregarding any prior knowledge or background reading, does the pattern observable in the datasets suggest the presence of any particular associations between pump locations and cholera cases?

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.

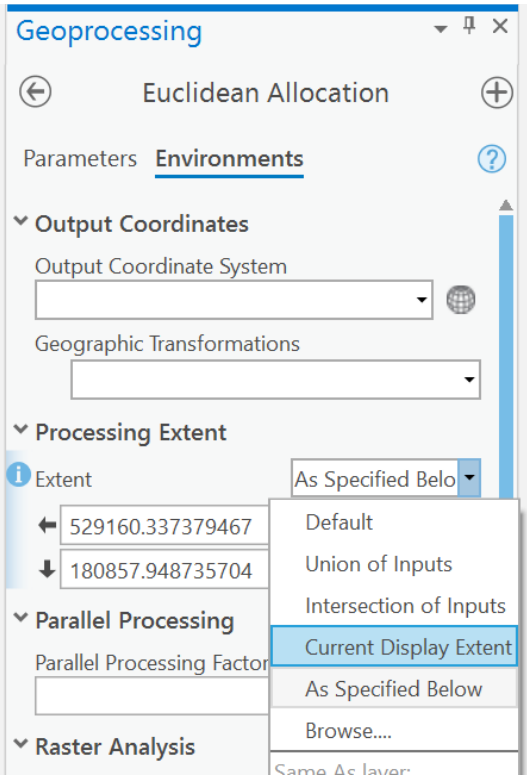
If you are new to ArcPro, you will need to make sure that you have the licence activated for the relevant extension (or plug-in). In this case, we will use *Spatial analyst*, ESRI's plug-in for raster data. To activate this, head for the *project* menu, then *licensing,* and then the *configure your licencing options* button. Activate the *spatial analyst* licence and use the top-left arrow icon to return (note the other extensions here: several others also have potential application in public health with for example business analyst having potential use in planning health services delivery).

Having activated this extension's licence, head for the *Analysis* menu and click on the *tools* button on the ribbon. In the right-hand 'geoprocessing' panel, use the search box to search for 'Euclidean', then run the *Euclidean Allocation* tool. Use the Help documentation to read about the *Euclidean Allocation* function. Run *Euclidean Allocation* based on the **Snowpumps** layer ensure that **Snowid or FID** is selected as the source field. See below for other suggested settings:



[A note on output file names: I often use geotiffs for raster output layers, a widely used raster format: specifying .tif when saving a raster to a folder will ensure it is saved in this format. In ArcPro, geotiff files have less stringent requirements in terms of allowable file and path names compared to rasters in geodatabases. It can only be used outside of geodatabases.]

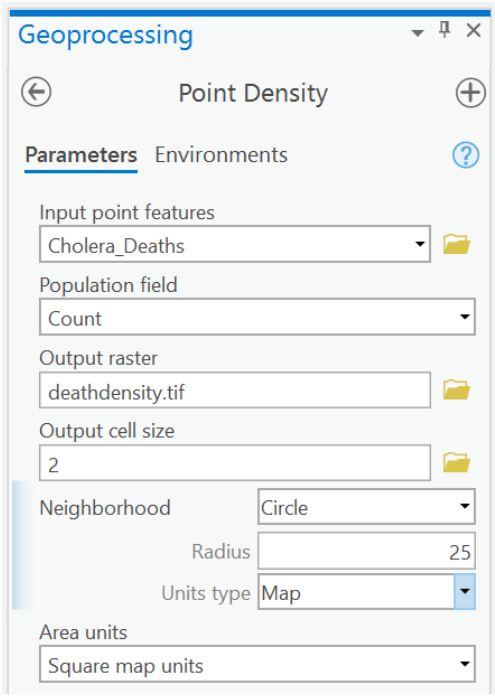
One more point – it can sometimes be worth changing the extent of an output raster. To do this, click on *environments* at the top of the tool dialog box, where you can then set the *processing extent*. This can be used to give the output raster the same extent as your map display, or one of your existing map layers:



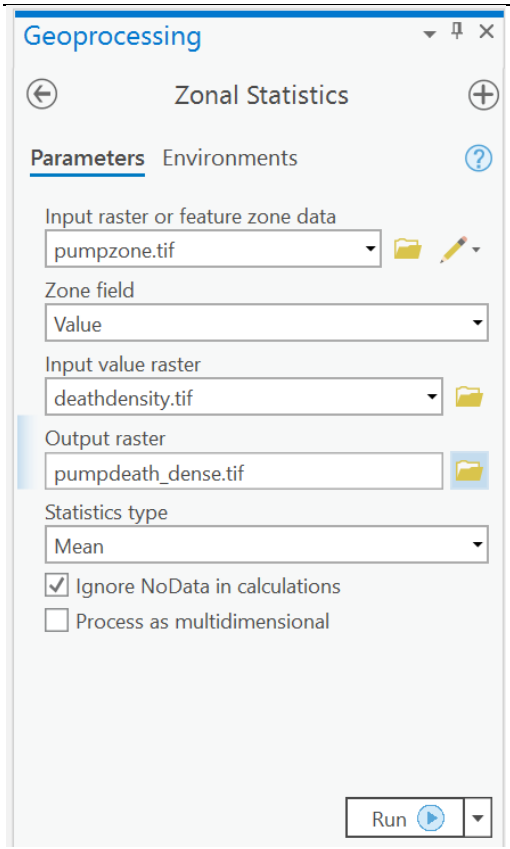
The resulting raster shows grid cell locations that are closer to each pump than to any other, storing the ID

of the nearest pump in the grid cell. This type of representation is frequently referred to as a Thiessen or Voroni tessellation and provides a rough idea of the area served by each pump. If you prefer vector to raster analyses, note that the tool *Create Thiessen polygons* will create very similar output, but in vector format. Now display the **Snowdeaths** layer over your new allocation raster.

Use the back button to return to the main geoprocessing panel. Then search for the 'point density' tool. 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). Densities (of deaths) will be calculated per unit area, using the measurement units specified at the foot of the dialog box. Explore the effect of using different search radii.



The tool *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 **Snowpumps** Euclidean allocation zones that you created earlier**.**

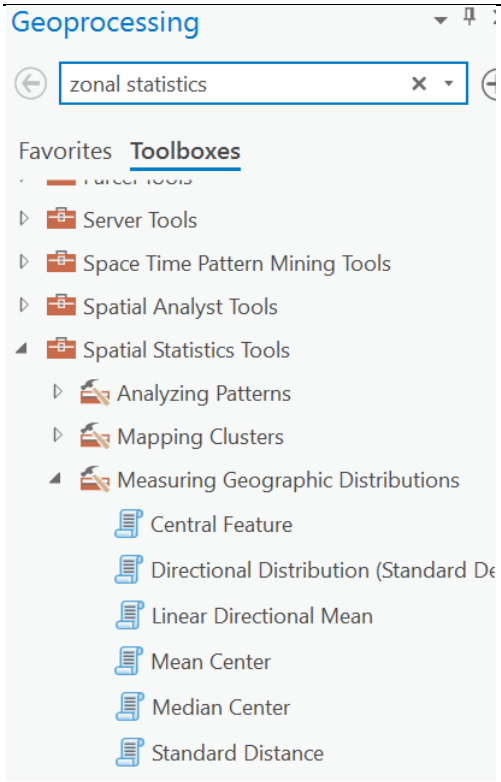


[Note: If you experimented with the tool *Create Thiessen polygons* earlier and wish to do a similar vector- rather than raster-based analysis, you could try using a tool such as *spatial join* to work out the number of points / deaths in each Thiessen polygon surrounding the water pumps]

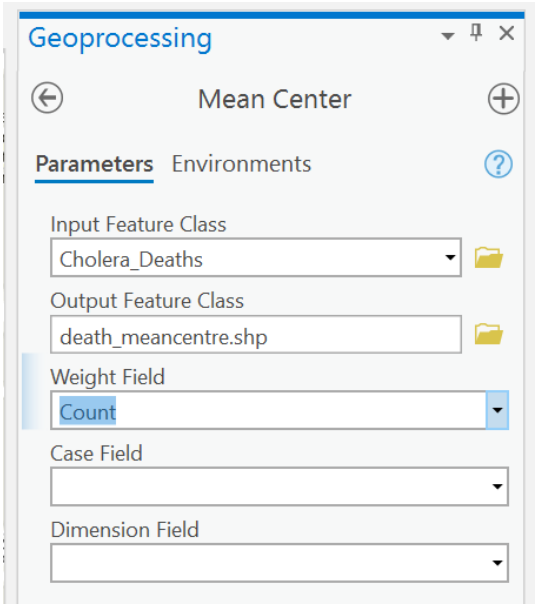
2. How would you interpret the results from the *Zonal Statistics* command? Do the results of *Zonal Statistics* confirm the association apparent from simply eye-balling the data?

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. Many epidemiological papers begin with a table of descriptive statistics (e.g. means and standard deviations of participants weights or ages). Somewhat analogously, it is possible to generate such simple descriptive statistics for the spatial component of data to gain an initial understanding of spatial patterns. To explore these, within the 'geoprocessing' panel, click on the *toolboxes* link at the top, which provides an alternative means of exploring the software's functionality. Head for *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* (note: weighted points, e.g. those addresses with more deaths, can 'count' for more via the weight field. Although we will not use it here, an optional *case field* can be used to generate different mean centre points, for example for cases in different years or months:



[Note: I have added .shp to my output file to save it in a folder in shape file format. You may wish to save yours in a geodatabase instead]

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*.

Try running the *directional distribution* tool too – you should find that this generates an ellipse that covers approximately two thirds of the deaths (the two axes of the ellipse reflect the standard deviations of distances of each point from the pattern's mean centre). However, the fraction covered by the ellipse in

practice is heavily dependent on the spatial pattern in the map layer.
Although these tools are simple, they are nonetheless very useful. Some examples:

- If you had an attribute field recording different periods when someone became infected or first diagnosed with a disease, you could use the *directional distribution* tool with these periods set as the 'case field' to generate different ellipses for different periods. You could then see if infectious disease cases remained concentrated in a small area or were spreading more widely.
- Similarly, if you had points representing the homes of people experiencing a given health outcome (such as a disease), and an attribute field indicating the health facility that they visited for treatment or diagnosis, you could use this tool with the health facility field as the 'case field' to generate simple, ellipse-shaped catchment areas for each facility.

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. Having reached the end of this preliminary spatial analysis, in what ways do you think the representation of the data affects the analysis? What assumptions underpin the analysis?

See next page for some comments on this.

Reflection exercises

2. Broadly yes – there appears to be a higher density of deaths around the Broad St pump. However, we cannot infer if the greater density of deaths around the Broad St pump is significantly higher than elsewhere.
3. These tools are purely descriptive of the spatial pattern in the deaths (there is no attempt to make any inferences about cholera risk from the pattern), but the mean centre and central feature do appear close to the pump.
4. Here are just a few examples of issues affecting the analysis:

- The data lack any representation of the spatial distribution of the underlying population at risk. Were there simply more people where there were cholera deaths reported, or were there more vulnerable people there? Controlling for the underlying distribution of the population at risk is often critical in spatial analysis, given that population is often spatially concentrated. Here, we implicitly assume a spatially uniform distribution of the population at risk.
- When examining deaths within Euclidean allocation zones, we assume that people fetch water from the pump closest to where they live.
- Does the location at which deaths are recorded reflect people's daily movements and their exposure to the *Vibrio* bacteria that cause cholera?