Compiling spatial data for a case-control study of Coronary Heart Disease

Overview:

In many parts of the world, a major challenge is that healthcare facility catchments often have very different boundaries to those used for reporting population data. Using data on disease cases reported at healthcare facilities can be particularly challenging without an accompanying understanding of the catchment population. The following exercise illustrates some of the difficulties involved.

A further challenge is understanding disease patterns across so-called **small areas**. Small areas are simply administrative units with small populations, typically containing 5,000 or less people. Examples of small areas in a US context might be **census tracts** (which typically contain around 4,000 people), as well as the smaller **census block groups** and **census blocks** that are the 'building bricks' that make up the larger census tracts. In a UK context, the equivalent to a census block is called a census **Output Area**, and the equivalent of a census block (or block group) is known as a **Super Output Area**.

In GIS and health analysis, calculating the number of disease cases for small areas is important because it can tell us about the likely caseload for a health facility (e.g. a new hospital or doctor's surgery). It can also help in health promotion campaigns, such as planning mailshots to specific postal or zip codes.

This exercise is concerned with estimating prevalence rates for small areas in a part of the UK.

Scenario:

The aim of this exercise is to calculate expected numbers of Coronary Heart Disease (CHD) cases for some general practices in the Cardiff area of the UK. In the UK, a general practice is a health facility offering primary care in the community and forms the first point of contact for patients in the UK healthcare system. By comparing the expected number of disease cases with the number of cases recorded as being treated on each surgery's computer, we may be able to identify areas where there are people with CHD who have not come forward for treatment.
Note that this exercise is intended to illustrate the difficulties of linking cases reported at facilities with population census data. It should be noted that the

exercise is rather unrealistic: for example, real GP practice boundaries would be very different to those used for this exercise, as catchments would vary in size and overlap somewhat. Similarly, the population data are based on an experimental data set from 2001, so the practical is intended to illustrate techniques, rather than provide a realistic depiction of rates.

Data files:

CardiffBay_census: This is a polygon shape file, containing synthetic census data for the Cardiff Bay area about housing characteristics and population. The polygons represent census Output Areas, each containing about 300 people. You can find out about Output Areas here:

https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography

The data we are using here are actually a prototype data set, used when a team at the University of Southampton were working to construct such small area boundaries.

Cardiff_censusoutline: an outline polygon showing the extent of available census data for the study site (Note: this has been prepared using the ArcGIS *dissolve* command from the **CardiffBay_census** map layer above).

chd_survey_results A CHD summary table from originally a sample of general practices across England and Wales, derived from:
http://www.heartstats.org/datapage.asp?id=1584

[This link is now broken, but similar statistics are available here:
https://www.bhf.org.uk/~/media/files/publications/research/2012_chd_statistics_compendium.pdf.]

generalpractices: This is a point shape file, which contains the locations of general practices in the Cardiff area of the UK. These locations were geocoded, based on postal (zip) codes. The data were recorded through a system for assessing the quality of care in general practice called the Quality and Outcomes Framework. The original link to the data is now dead, but here is one of several places from which you can download such data:
https://www.gpcontract.co.uk/download.
As well as the zip / postal code and grid reference for the practice, the table of attributes contains a field called **qofcases**, which contains the number of CHD cases as recorded on each practice's computer.

practice_bnds: This is a polygon shape file of the boundaries of each practice, which was produced using the ArcView *Euclidean Allocation* function. It assumes that each Cardiff resident will go to their nearest general practice.

A note about map projections: These data sets are in a local geographical reference system widely used in the UK known as the British National Grid (or sometimes Ordnance Survey National Grid). This reference system records locations in metres relative to an origin point in the sea, just off to the southwest of the British Isles. The National Grid is very similar to the Universal Transverse Mercator co-ordinate system, used in many other parts of the world. You may receive some warning messages in a green font from ArcView about map projections. Do not be too alarmed about these non-fatal warning messages!

Data Attribution:

Practical Exercise:

Begin by viewing the data that you have been provided with in ArcGIS Pro and familiarise yourself with the three map layers. You will need to create a new project and connect to your working folder as you did in previous practicals.

Work out expected numbers of cases for CHD by output area

Our first task is to work out how many people are likely to have CHD in each of our 'small areas', the output areas in the **CardiffBay_census** map layer. The spreadsheet called **chd_survey_results** contains a summary table from a national database of a sample of general practices across England and Wales. The summary table shows CHD prevalence rates, broken down by age, sex and levels of deprivation.

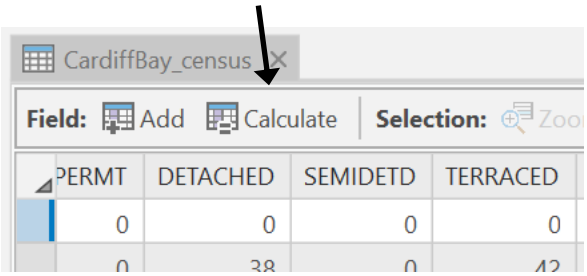
Task 1. Take a look at this spreadsheet. Do you consider age, sex and deprivation to be predisposing, individual, or environmental factors? How do they affect CHD, according to this table?

We can use this table to calculate a very simple expected number of CHD cases for each of our small areas, based on age. A simple average of the figures in this

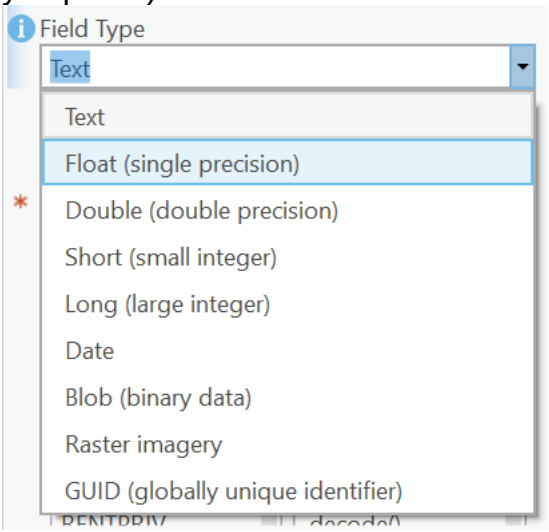
table (see row 27 of the spreadsheet) suggests that those aged over 65 years have much higher rates of CHD nationally – 17% have CHD compared to just 2% in those under 65 years.

Open up the attributes of the **CardiffBay_census** map layer (right-click on the name of the map layer in the left-hand layers panel and then choose *attribute table*).

You can use *calculate* to work out the population under 65 years in each area.

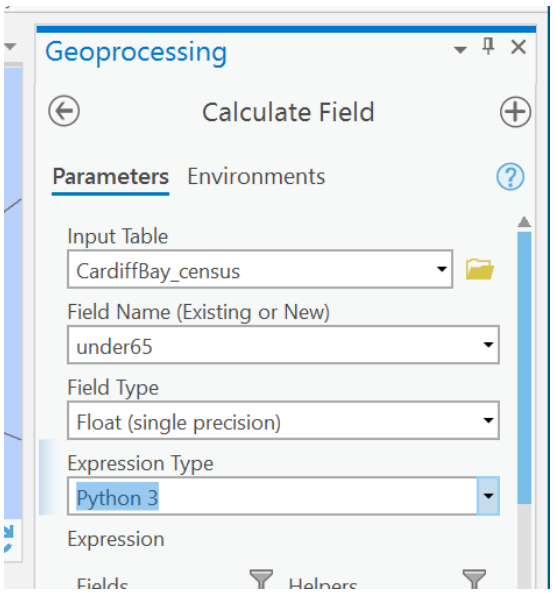


Calculate will add the results of a calculation to a new or existing field. In this case, we will use it to create a new field called **under65** of type *float* (or *integer* if you prefer):



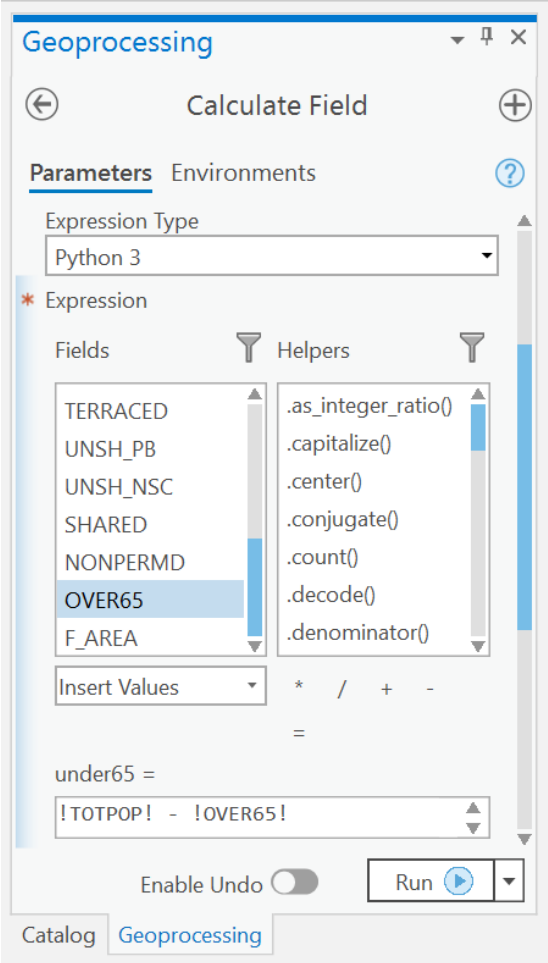
[A note on the more obscure field types: short and long integer fields hold whole numbers; float and double fields hold numbers with fractional parts; a GUID is an identifier that is unique to a particular record and does not appear anywhere else in a geodatabase; whilst a blob field holds binary large objects, which could be pictures or entire files]

Leave *Expression type* set to **python 3** [again, a word of explanation: Arcade is a coding syntax specific to ArcGIS and ESRI, designed to be easy to use and to handle spatial data, whilst Python is a widely used coding language, also used for example within QGIS and in mathematics in the form of NumPy]**:**



In our attributes, we already have a field with the total population of each output area (**totpop**) and a field with a count of those aged over 65 years in the population (**over65**). We can use these two fields to calculate our **under65** field as follows:

- Double-click **totpop** in the *fields* list
- Select '-'
- Double click **over65:**



With Python, the names of fields are enclosed by exclamation marks, so the expression box follows Python syntax for subtracting the number of people over 65 years from the population in each area.

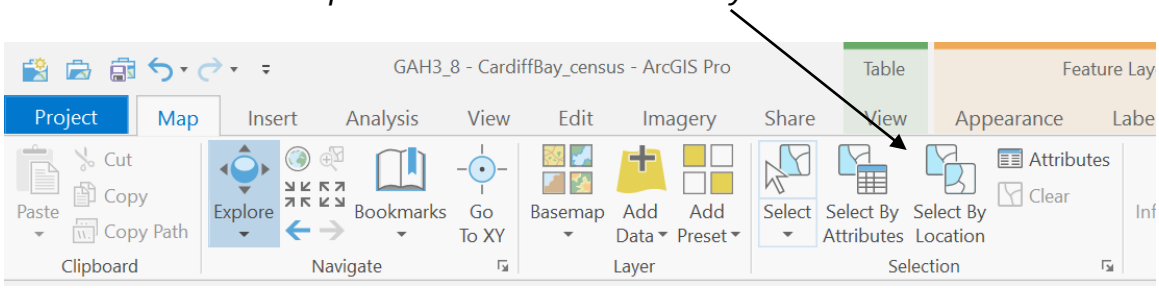
Task 2: Using the *field calculator* menu option (as we've just done) and the information in the spreadsheet **chd_survey_results,** try estimating the number of CHD cases among those aged under 65 years and over 65 years respectively. You will need to assume that the national CHD rates in these areas hold true for each of our small areas (see bottom of p. 3). Add the number of cases in each group together to produce a count of total CHD cases for each of our output areas – let us call this **chdcases**.

Find the practices that are completely within the area with census data

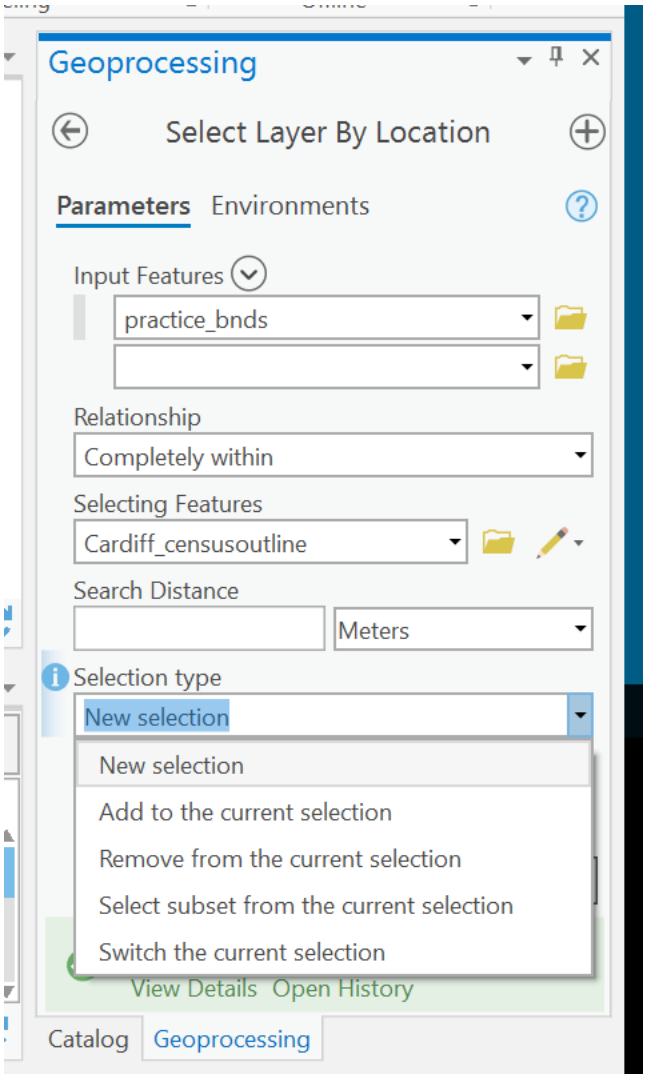
We have an initial problem to solve first of all, which is that our two sets of boundaries do not have the same extent. Some of our practice catchment boundaries extend well beyond the area of available census data.

To resolve this problem, we can do the following:

- Head for the *map* menu and choose *select by location*



- Next choose **practice_bnds** as the *input features* and **Cardiff_censusoutline** as the *selecting features*.
- Set the *relationship* to be **completely within**. This will mean only practice boundaries that are completely within the area included in the output area map will be selected.
- As the *selection type*, choose **new selection** (other options here, such as *subset from current selection* will take account of any practices that are already selected, only selecting practices that are already selected and within the output area map layer too):



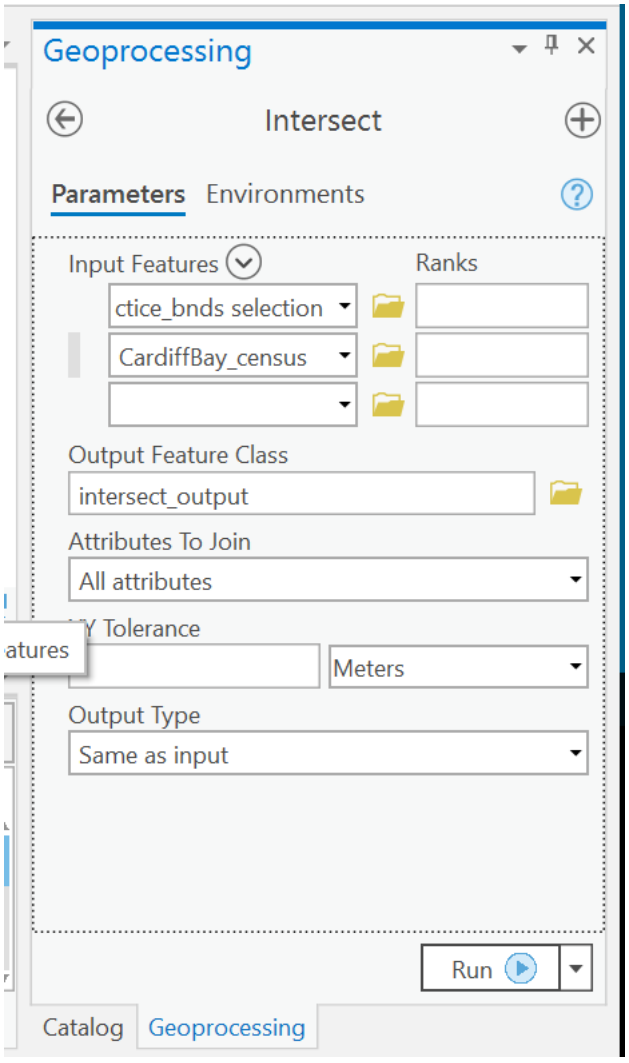
When you run the tool, you should find that only the practice boundaries completely covered by output areas have been selected. You can save these as a new map layer by right-clicking on the left-hand layers panel, the choosing *selection* and *make layer from selected features*.

Work out expected numbers of cases for CHD by practice

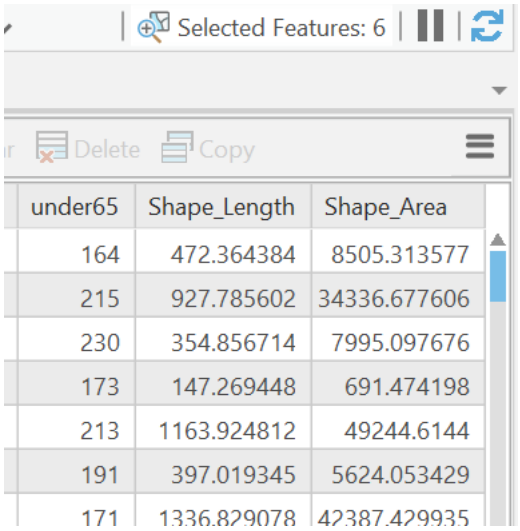
We now have a map of the expected number of cases of CHD for each of our output areas, but the data on cases being treated in Cardiff relate to general practices, which have different boundaries. We now need to calculate the expected number of CHD cases for each general practice catchment area, rather than by census output area. This is what is known as an **areal interpolation** problem – when we need to take attributes for one polygon map layer and transfer them across to a second map layer with different polygon boundaries.

This is quite a complex process! We first have to intersect the practice and census boundaries, then work out the proportion of CHD cases in each fragment of a practice catchment. Assuming an even distribution of population, we can divide up the cases according to area (so for example a census area that's split into two fragments of equal area with have equal numbers of CHD cases in each half).

To undertake this process, the first step here is to intersect our set of practice boundaries with complete census data (**practice_bnds selection** created above) with our census data, **CardiffBay_census**. To do this, head for the geoprocessing panel, select *analysis tools* and then *overlay / intersect*, creating a new file called **intersect_output** or similar (note: we recommend storing the output in a geodatabase, rather than as a shape file, as will be explained later*:*



If you run the *intersect* tool and open up the attribute table of your output layer, you should find that the table contains the attributes from both the output areas and the practice boundaries:



under65	Shape_Length	Shape_Area
164	472.364384	8505.313577
215	927.785602	34336.677606
230	354.856714	7995.097676
173	147.269448	691.474198
213	1163.924812	49244.6144
191	397.019345	5624.053429
171	1336.829078	42387.429935

Notice also that your layer should have two automatically calculated fields (because it is in a geodatabase) called *shape_length* and *shape_area*. These contain the perimeters and areas of the intersected polygons (based on overlaying the two input layers) in metres / metres square.

Now we can use this area field for our newly created polygons to finalise our areal interpolation calculation:

- In the new layer's attribute table, use *calculate* to divide **inter_area** by **f_area**, placing the result in a new field of type *double* called **prop_area**. This tells us what proportion of the original input layer polygon is occupied by the new, sub-divided polygons in the output layer.
- Again, using *calculate,* multiply **prop_area** by **chdcases** (this contains the total expected cases from Task 2 above) and place the result in a new field **interchd** of type *double*.
- Finally, right-click on the header of the **practID** field and select *summarise*. Choose **practid** under *case field* and choose to calculate the *sum* (under *statistic type*) of **interchd** as the *statistics field* that you just created. Under *output table*, store the result as a new table **prac_chd.** *Input table* should already be specified.
- We now have a table that has the number of CHD cases for each practice…finally! To map this out, we need to join this back to our practice boundaries file. To do this, close down any attribute windows you may have open and then right-click on the **practice_bnds selection** layer. Select *joins and relates* and then *add join*. Under *input join field* select **practid**. Under *join table* select **prac_chd** and under *output join field* select **Practid** and choose Run**.** ArcGIS will now match up entries for

PractID in the **practice_within** map layer with those in the **Prac_chd** table.

- If you look at the **practice_within** layer now, you should find that the information on chd cases has been added to its table of attributes.
- Produce a thematic map of the resultant expected CHD cases per practice catchment.

Of course, there are more sophisticated ways of tackling this spatial interpolation problem. For example, if we had some idea of the location of the population within each of the census areas (such as the locations of individual post or zip codes for example), we could use this to information to help interpolate the census information into practice boundaries.

Extension exercise: Calculate standardised rates of CHD

Task 3: By manipulating the table of attributes of the **practice_within** map layer, work out a Standardised Morbidity Rate (SMR) for CHD for each practice

(Hint To do this, you will need to divide the observed number of CHD cases [i.e. the actual number of CHD on the practice's computer, stored in the field **QOFcases**] by the expected number, given each practice's population [that you have just calculated]. Produce a map of the resultant SMRs.

Assess how well you think the standardisation of CHD disease cases has worked. What aspects of our data or GIS analysis do you think might have influenced the final map of standardised rates of CHD? Can you think of any ways that they might be improved?