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

Overview:

In many parts of the world, a major challenge is to produce estimates of the number of disease cases for 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. The sections of the exercise that appear in boxes form the basis for your assignment.

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.

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. Further details are available via the output area demonstrator project.

http://www.public.geog.soton.ac.uk/research/oa2001/oademon.asp

The data we are using here are taken from the Cardiff test area for **scenario 7**. Note that there is a description of the attributes of each output area available via the link at the top of the Output Area Demonstrator Project page (in paragraph 2).

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

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 - see
http://www.wales.nhs.uk/sites3/page.cfm?orgid=480&pid=10486
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:

http://www.ordnancesurvey.co.uk/oswebsite/opendata/docs/os-opendata-licence.pdf

Practical Exercise:

Begin by viewing the data that you have been provided with in ArcMap and familiarise yourself with the three map layers.

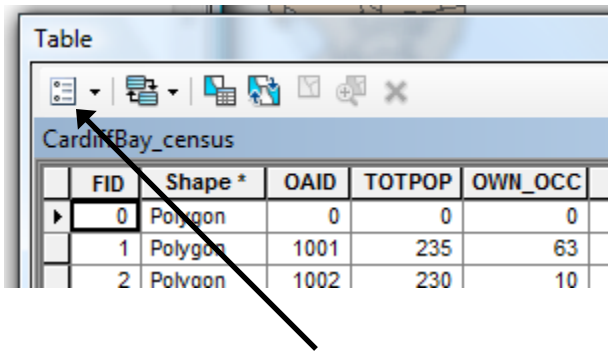
Work out expected numbers of cases for CHD by output area

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 and then choose *open attributes*).

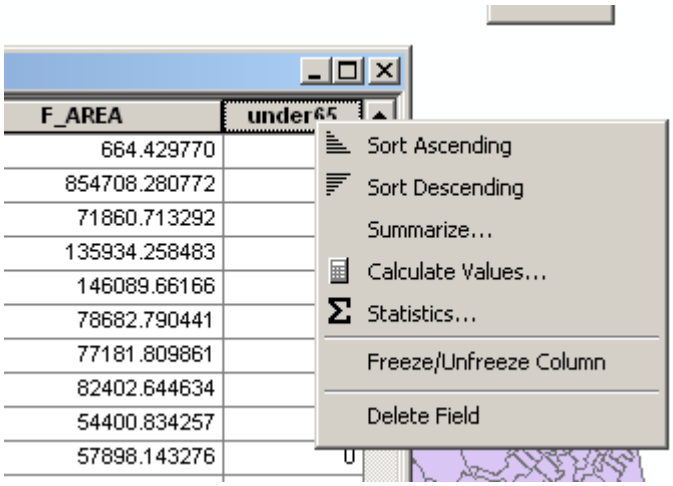


By clicking on the *table options* button, use *add field* to add in four new fields to this attribute table, all of type *long integer*:

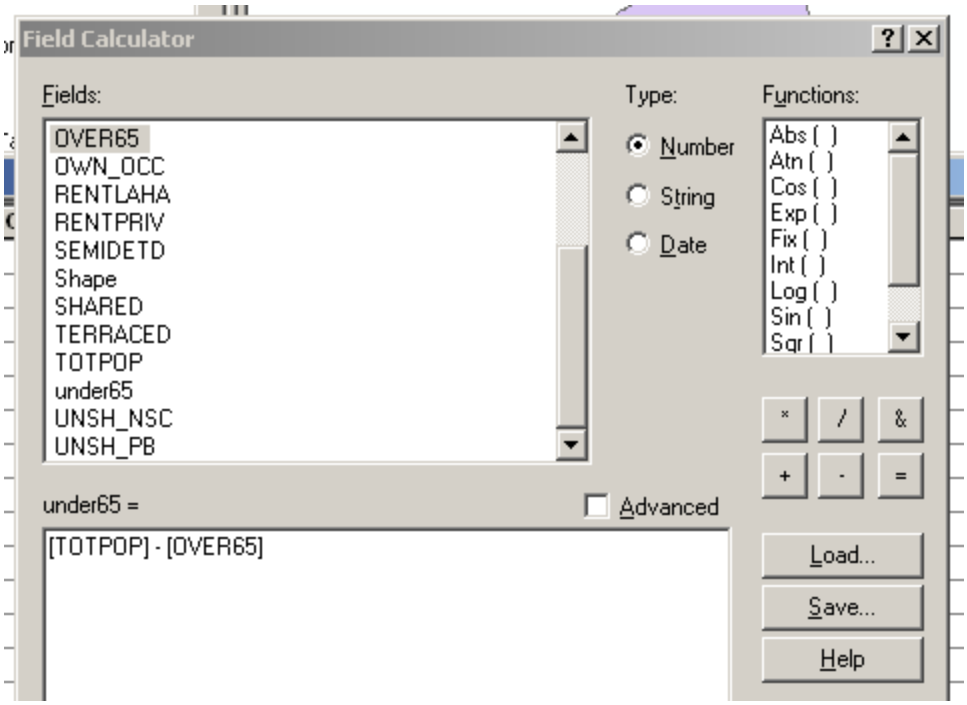
- a new field for the total population aged under 65 years – call this field **under65**;
- a new field for the estimated number of CHD cases among those aged under 65 years;
- a new field for the estimated number of CHD cases among those aged over 65 years;
- a new field for the total estimated number of CHD cases in any age group

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:

- Still looking at the table of attributes, right-click on the header of your new **under65** field and choose *field calculator.* Choose '*yes'* to when prompted by the warning message about doing calculations outside of an edit session and leave the *parser* set to *VB Script.*



- Next, in the left-hand part of the 'calculate' screen, click on **totpop**, then on the minus ('-') button below 'functions' and then click on **over65** back in the left-hand part of the screen. This will subtract the number of people over 65 from the total population, to give us those aged under 65 years.



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. 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**. Create a map of the result and include this in your report.

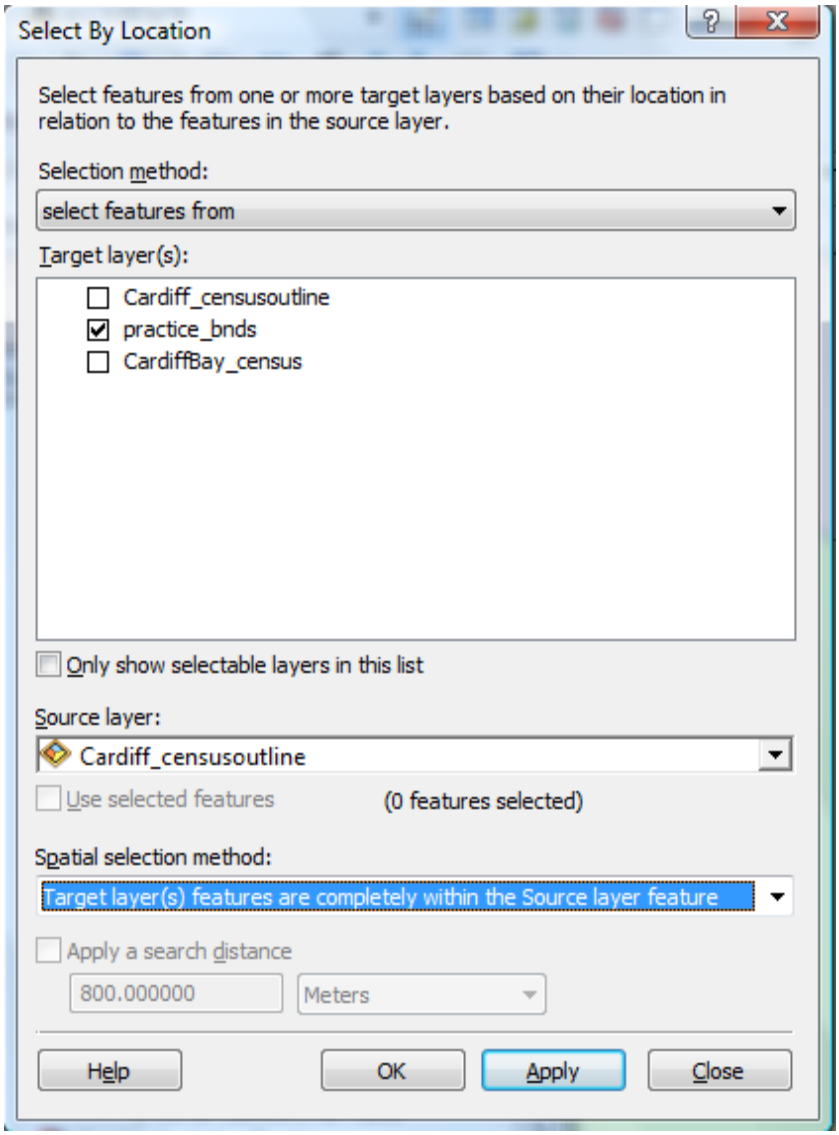
Find the practices that are completely within the area with census data

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 *selection* menu and select *select by location.*
- Leave *select features from* as the *selection method* and set **practice_bnds** as the *target layer.* Select **Cardiff_censusoutline** as the *source layer* and select *target layer features are completely within the source layer feature* (see below)

- Right-click on this map layer, choose *data* and then *export data*. If we choose here to *export…selected features* we can generate a new layer just with those catchments that are entirely contained within the available census data. Call this **practice_within** or something similar.

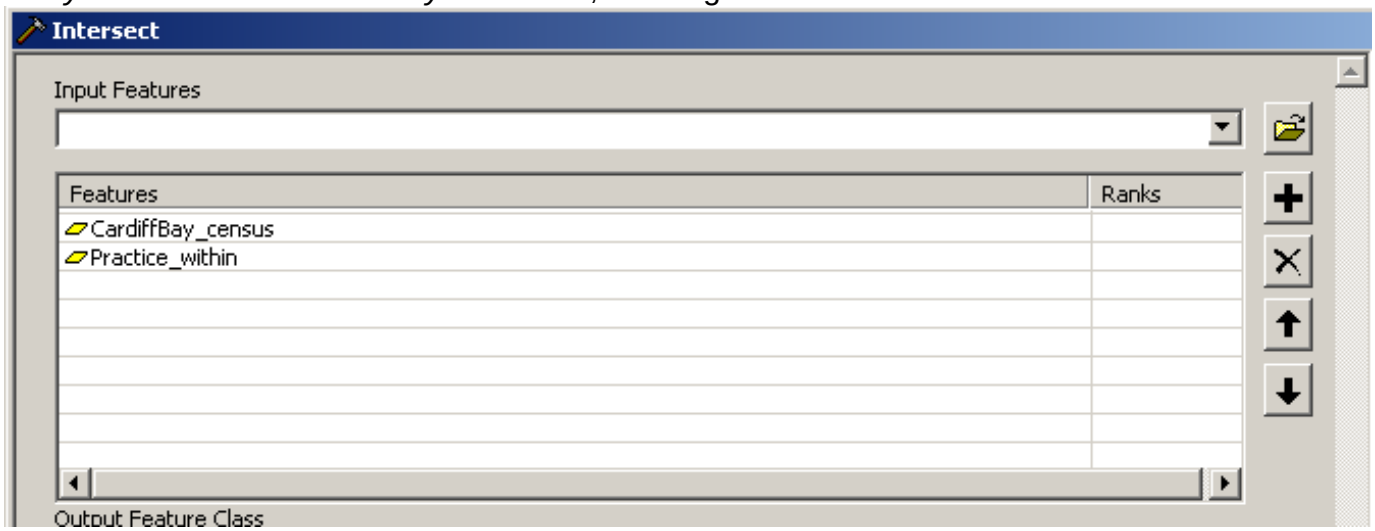


Work out expected numbers of cases for CHD by practice

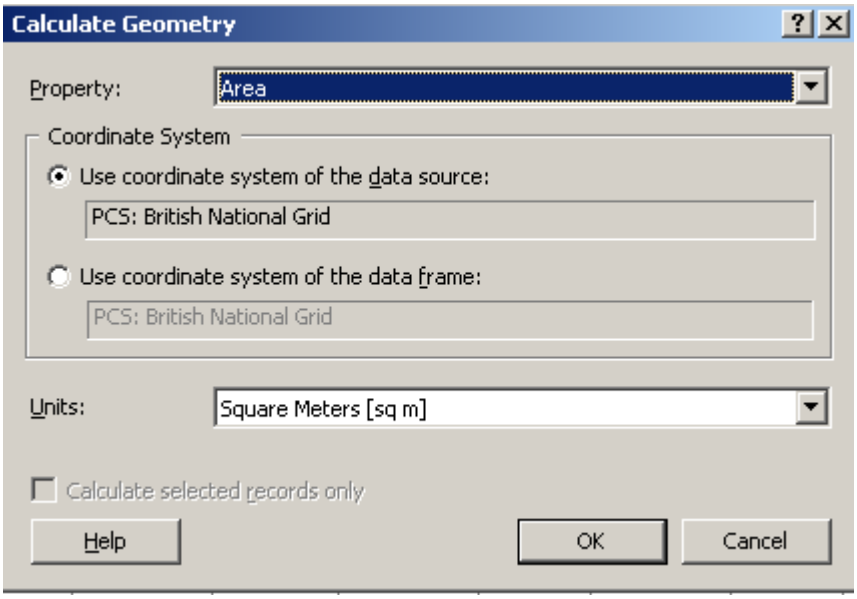
We now have a map of the expected number of cases of CHD for each of our output areas, but the data on cases being treated in Cardiff relate to general practices, which have different boundaries. We now need to calculate the expected number of CHD cases for each general practice catchment area, rather 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_within** created above) with our census data, **CardiffBay_census**. To do this, head for the ArcToolBox, select *analysis tools* and then *overlay / intersect*, creating a new file called **intersect***:*



- We now need to open up the attribute table of our new **intersect** map layer and create a new field of type **double** called **inter_area.** This will hold the areas of each new polygon that has been created.
- Whilst looking at the attributes of this table, right-click on the header of this new field **inter_area** and select **calculate geometry.** Click yes to any error messages you might see. Make sure **Area** is selected as the property to calculate and choose *ok.*



- create another new field, again of type *double* called **Prop_area**
- Use the *field calculator* option again (accessible by right-clicking on the header of this new field **prop_area**) to divide **inter_area** by **f_area**, placing the result in this new field.
- create yet another new field of type *double* called **interchd**
- Again using the *field calculator*, multiple **prop_area** by **chdcases** (this contains the total expected cases from Task 2 above) and place the result in this new field **interchd**
- Finally, right-click on the header of the **practID** field and select *summarise*. Choose **practid** under *select a field to summarize* and choose to calculate the *sum* of the **interchd** field that you just created. Under *specify output table*, store the result as a new table **prac_chd.**
- 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_within** layer. Select *joins and relates* and then *join*. Choose to *join attributes from a table* and under *choose the field in this layer that the join will be based on* select **practid**. Under *choose the table to join to this layer..* select **prac_chd** and under *select the field in this table to base the join on* select **Practid** and choose OK**.** 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.

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 and include this in your report.

Task 4: 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?