

## Practical: Geocoding data using ArcGIS

### Overview:

This exercise looks at how to use ArcGIS geocoding tools to create a vector point map layer from information about postal or zip codes. In particular, we use post code data on general practice surgeries in the UK.

### Data:

#### Data attribution:

- Contains Ordnance Survey Data © Crown copyright and database right 2011.
- Contains Royal Mail Data © Royal Mail copyright and database right 2011.

#### Map layers and tables for the exercise:

The data for this exercise are taken from two sources:

- **CodePoint.csv** - This table is derived from the Ordnance Survey's Open Data initiative (<http://www.ordnancesurvey.co.uk/oswebsite/opendata/>), and in particular a data product called CodePoint® Open. This contains X and Y coordinates for unit postcodes. In this context, a unit postcode is an alphanumeric code (e.g. SO17 1BJ) that is associated the location of the most centrally located property within small groups (typically 10 to 30) of properties in the UK. The postcodes are used for both residential and non-residential properties. In this case, we have extracted the CodePoint® Open data for the Cardiff post code area. The table contains three fields - the **postcode** and its **xcoord** and **ycoord**, expressed in metres on the British National Grid coordinate system.
- **Surgeries** This is a table that contains the postal (zip codes) of practices in Cardiff (in the **zipcode** field) and reported cases of coronary heart disease (in the **qofcases** field) at each one. The table comes from the Quality and Outcomes Framework for primary care, available here: <http://www.wales.nhs.uk/sites3/page.cfm?orgid=480&pid=10486>

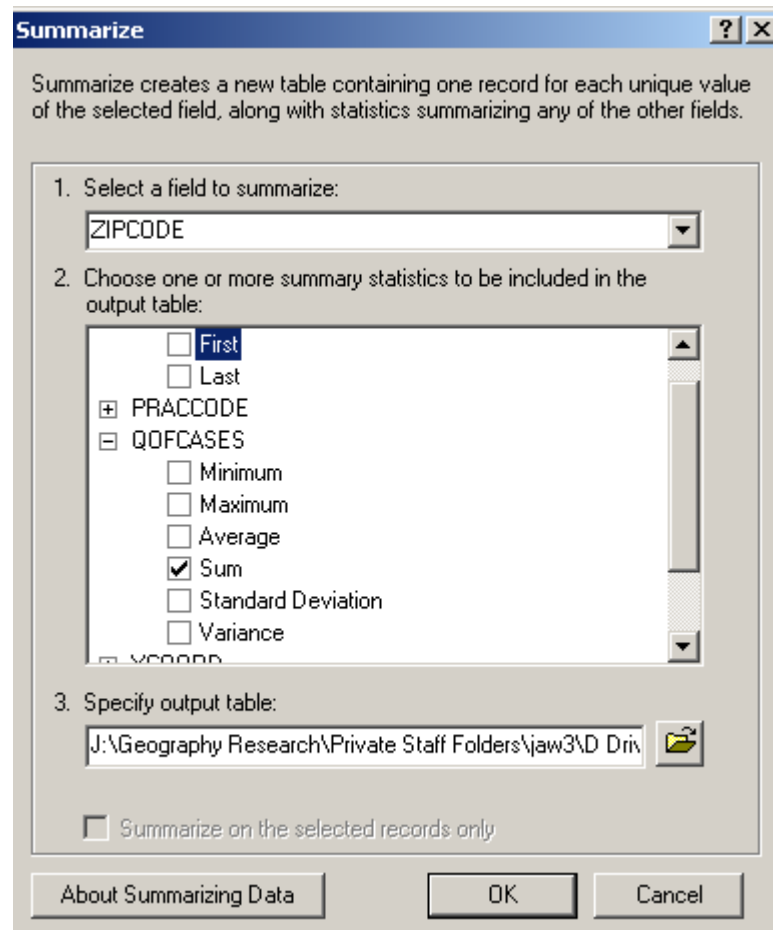
In an international context, note that reference data for geocoding - including the post codes used here - can also be downloaded from the GeoNames web site (<http://www.geonames.org/>). This site includes data resources for other countries, not just the UK.

## Practical Instructions:

### Prepare the general practice data:

Open up the **surgeries** data file in ArcMap. Once you have loaded it up, right-click on the name of this map layer in the left-hand table of contents panel and choose *open*. Sort the data file by zip (postal) code, by right-clicking on the **zipcode** field and choosing *sort ascending*. You will see that there are some duplicate entries in the field - for example there are two entries for CF10 5UZ, resulting from the way the data have been compiled. We will need to amalgamate these entries before processing the data further.

To resolve this problem, right-click on the **zipcode** field and select *summarize*.



For each zip code, you should then be able to generate a *sum* of the number of reported heart disease cases, stored in the **qofcases** field. You will also need to choose an appropriate name for the output table, e.g. **practices**.

## Produce a point map layer of post code locations to use as reference data

Now open up the **Codepoint.csv** file within ArcMap. Right-click on this map layer in the left-hand table of contents and choose *open* and inspect its contents. You will see that there is a postcode with an associated X and Y coordinate in the British National Grid, a reference system specifically designed for Great Britain. Close down the table and map these points as follows:

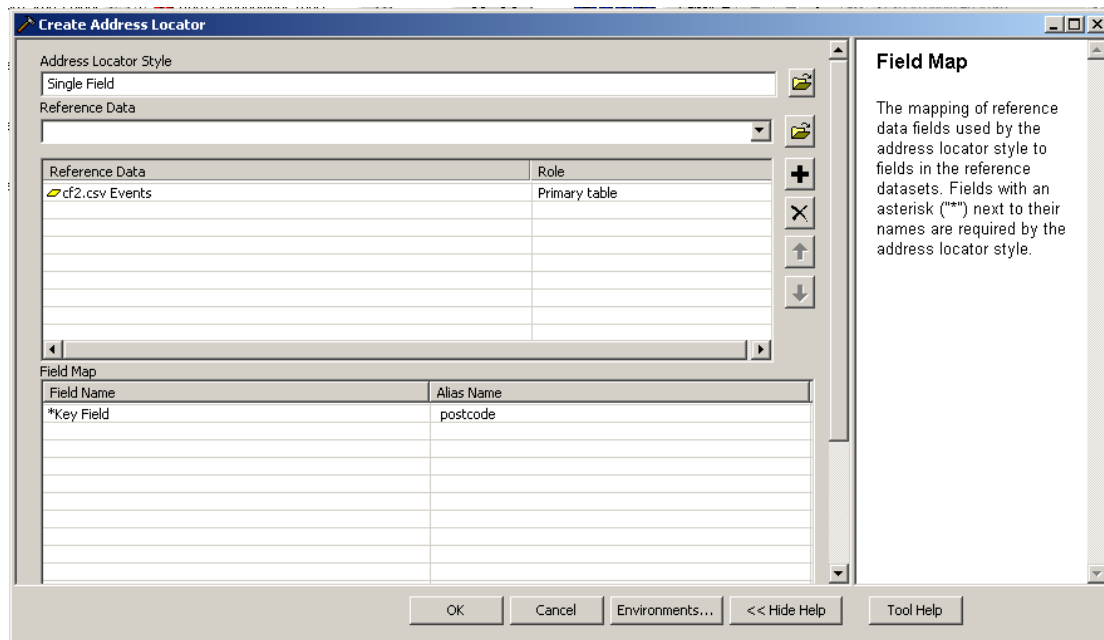
- select *add XY data* from the *add data* menu option on the *file* menu
- Make sure that the *X field* is set to **xcoord** and the *Y field* is set to **ycoord**.
- Next, click on the *edit* button and choose *select* to pick a pre-defined reference system. Select the *projected coordinate systems* folder, then the *national grids* folder, and then the *British National Grid* file within the *Europe* folder. Click on OK and OK again to map the unit postcodes. The city of Cardiff is the large group of postcodes in the southeast, whilst to the north, there is a mountainous area incised by valleys, where the settlement pattern and therefore postcodes follow the valleys.

Before using this information for geocoding, we need to convert it from events (i.e. where there is a 'live' link between the coordinate fields in the attribute table and the points on the map) to a shape file or geodatabase, where such a link is removed. To do this, right-click on the **codepoint events** in the left-hand table of contents and choose **data** and then **export data** and export the data to a shape file called **codepoint\_cardiff**.

## Create an Address Locator

In ArcGIS, an Address Locator is a set of processing instructions associated with a particular reference data set (a map layer that has placename, zip or postcode or address attribute fields associated with vector features). In order to geocode our data, we first need to create an address locator.

Go to the ArcToolBox, select Geocoding Tools, and then choose *Create Address Locator*. There are many different styles of address locator that can be developed - you can see the different types under the *address locator style* option on the dialog box. For this exercise, select the *general - single field* style, which is the simplest of these types.

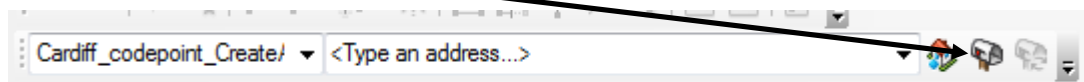


Next, we need to select some *reference data* - in other words, data which will provide us with the point locations of our practices. Select the **codepoint\_cardiff** map layer for this. Set its *role* to be **primary table** (note: it is possible to use other ancillary data here - for example where the same street is known by several different names, we can use an *alias table* here). In the *field map* at the bottom, we need to indicate which field in this **codepoint\_cardiff** attribute table has the postcode attributes in it. In this case, the field is **postcode**. Finally, at the bottom of the screen, choose a name for your new address locator. Click OK, and the software will create a new address locator for you.

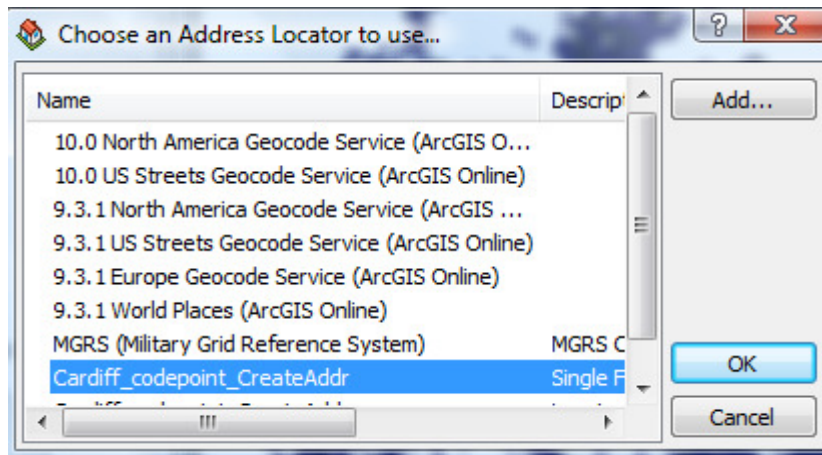
### Geocode the practice postcodes:

Next, right- click on the toolbars area at the top of your screen in ArcMap and then check the 'geocoding' toolbar, so that this appears on your screen.

You should see a set of icons like those shown below. Click on the 'geocode addresses' icon.

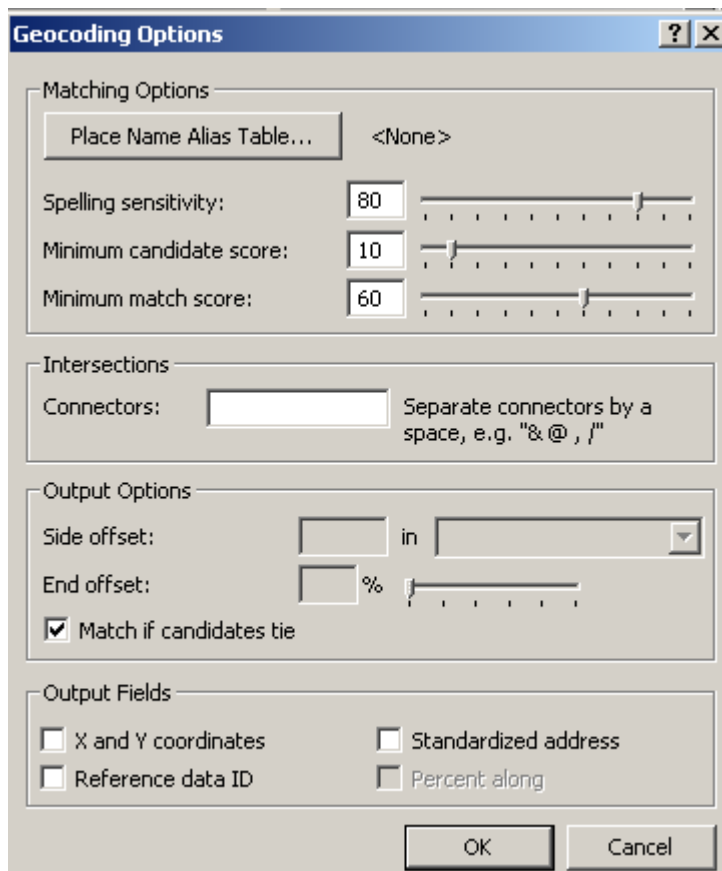


You will now be asked to *choose an address locator to use*. Select the address locator that we created earlier, which you should be able to navigate to by clicking on the *add...* button on this screen and click on OK.



On the next screen, you will need to select the *address table* or data to be geocoded - namely our **practices** file. The field that we will use to match to our map layer of postcodes is **zipcode**. Under *output shapefile or feature class*, we will also need to specify a name for our output, e.g. **geocoded\_practices**. Leave the default setting of *create static snapshot of table inside new feature class* selected.

Clicking on *geocoding options* provides some more control on this process:



- ArcGIS will generate a score between 0 and 100, indicating how closely entries in our **codepoint\_cardiff** postcode field and **practices** postcode field match. The *spelling sensitivity* setting enables us to

vary the strictness of the scoring system. In situations where spelling variation is more likely (e.g. where placenames have been translated into English say), we can reduce this sensitivity parameter to account for this.

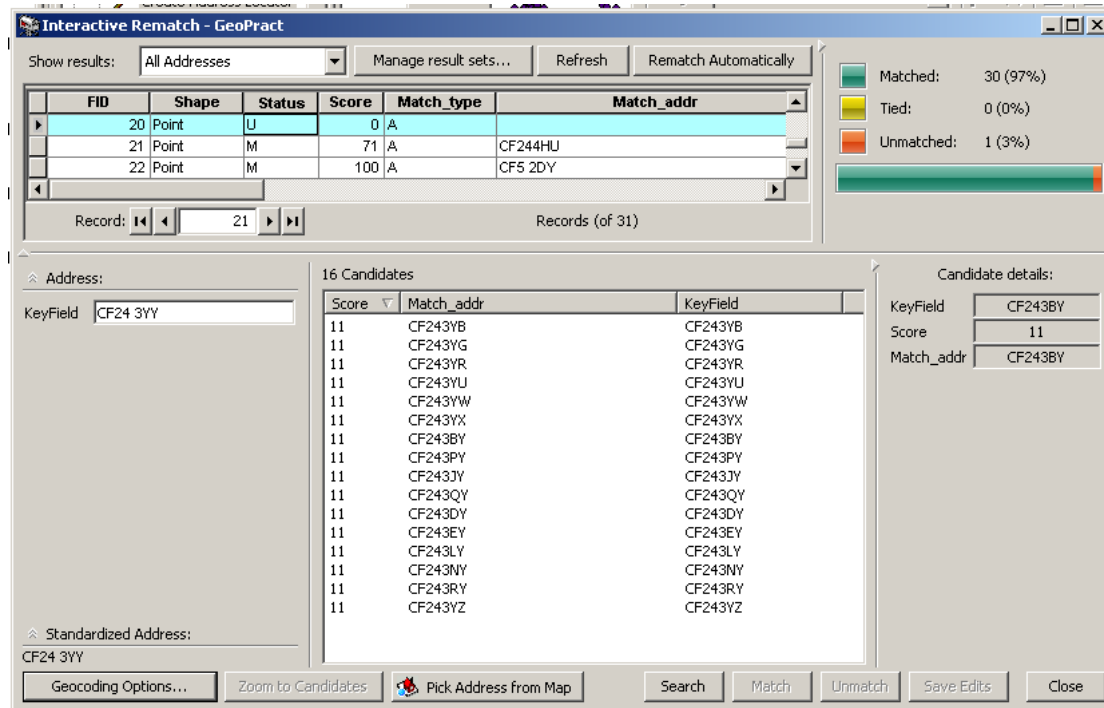
- The *minimum match score* is the lowest score that ArcGIS will accept as being close enough for an entry for a **codepoint\_cardiff** postcode and a **practices** postcode to match them together. If the score for word similarity is lower than this, then ArcGIS will not link our practice data to one of our point locations.
- The *minimum candidate score* helps ArcGIS handle matches that seem quite close, but remain below the *minimum match score*. If for example a possible match on a postcode or place-name scores 70 on a similarity scale of 0 to 100 (e.g. 'New Yor' rather than 'New York'), we may want to review this later and match it up interactively. Such 'near misses' are called *candidates*. ArcGIS will keep a list of possible candidates for each of our **practices** postcodes that can be matched up later interactively by the software user.
- In summary, in looking for possible matches in our **codepoint\_cardiff** layer, ArcGIS will produce three sets of results: *matches*, where there is a close or perfect match between entries in our two tables; *candidates*, where there is a weaker, potential match between entries that can be reviewed interactively later, and *unmatched* records.
- At the bottom of the screen, we have the option of including additional attributes, such as X and Y coordinates for the postcodes we successfully geocode.

For now, let us leave these settings at their default values, so press OK and continue.

ArcGIS will start to process the data and present a 'traffic light' system, indicating how much of the data has been geocoded (green indicates matches; red indicates unmatched records). We can either accept our results by pressing *close* at this point or else choose *rematch* to start matching up records interactively. Choose *rematch* here.

## Interactive geocoding

You will now be presented with a screen for manually reviewing and matching the post codes of surgeries. The top right of this screen shows a traffic light-coded view of progress in geocoding our surgery locations. 30 surgeries have been automatically matched (in green). There are none that are tied (i.e. where there are two or more postcodes in our **codepoint\_cardiff** data set that are equally close to a postcode in our **practices** data file), shown in yellow. There is 1 postcode that is unmatched, shown in red.



The table of information at the top of the screen tells us something about the status of each of our **practices** post codes. 'M' indicates a surgery postcode that has been successfully matched to a corresponding **Codepoint\_cardiff** postcode. 'U' indicates a postcode that remains unmatched to a corresponding postcode. The **Match\_type** indicates whether the match was made automatically by ArcGIS ('A') or manually through this screen ('U').

It is possible to sort the contents of this screen to make it easier to navigate. You can right-click on a field name (e.g. **status**) and select *sort ascending*. You can also use the *show results* dropdown to look at subsets of records, such as those that were matched but with scores below 80 for example.

If you scroll through the different postcodes for practices, you will see that the one unmatched practice postcode is for a practice with postcode CF24 3YY.

## What can we do to geocode any unmatched postcodes?

One option is to click *geocoding options* and then change the default parameters that are used to match up the names in our reference data and addresses that we wish to geocode:

- Click on *geocoding options*, and try setting the *spelling sensitivity* to be lower (e.g. to 60) and set the *minimum candidate score* to 10. This will mean that the score representing the degree of similarity between corresponding entries in the **codepoint\_cardiff** and **practices** data sets will be rescaled and the software will be more tolerant of mismatches in spelling (e.g. a one letter difference in

placenames between the reference data and addresses might result in a score of 90 rather than 80). Often, the spelling sensitivity parameter might be reduced where there were issues around names being translated from another language (e.g. Arabic into English). After changing such settings, you have the option to *rematch automatically*, applying these new settings to the whole data set.

- After adjusting the spelling sensitivity, there should now be a set of potential candidate matches for our unmatched post code. Each scores 11 out of 100 as a match and is therefore over the *minimum candidate score* of 10. We can take a look at where these potential matches are on the map (via *zoom to candidates*).
- If we identified a mistake in the left-hand *address keyfield*, we could type into this box and correct it, and then press *search* to resolve the problem as another alternative.
- Let us assume that we had a data entry here with an X being mistyped as a Y, and let us match our unmatched record CF24 3YY to postcode CF24 3YX.
- To do this, click on the CF24 3YX entry on the list of candidates and press *match*.
- We should now have a complete set of matches for all of our practice postcodes, so press *close*.