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 (https://osdatahub.os.uk/downloads/open), 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 four key fields – the **postcode** and its **xcoord** and **ycoord**, expressed in metres on the British National Grid coordinate system. **Newpostcode** is a variant of the postcode field, but with an extra space added, which we will use later.
- **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 originally (link now dead): 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. If you want a replacement link to NHS treatment facilities to replace the 'dead' link above, you could head here: https://digital.nhs.uk/services/organisation-data-service/data-downloads/miscellaneous

Note: if you have time, you could experiment yourself with some of the techniques in this practical using these other data resources.

Practical Instructions:

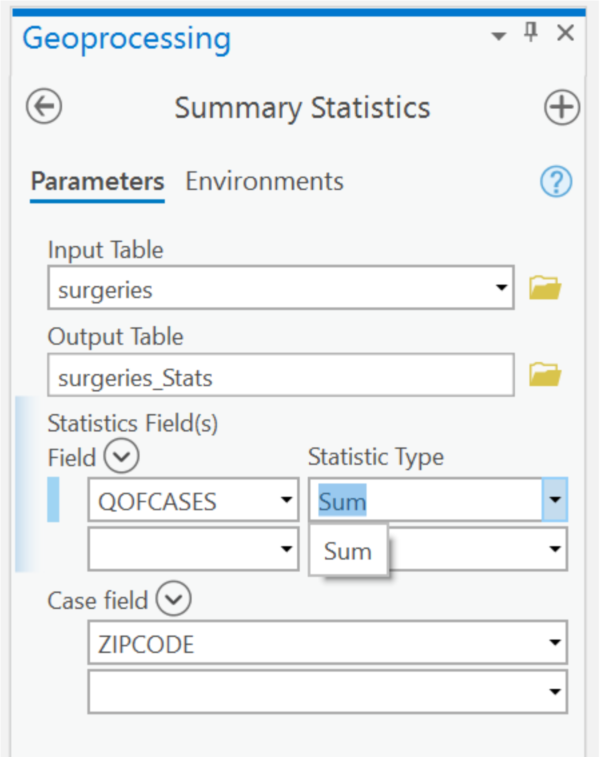
Starting ArcGIS Pro

Download the data for this exercise into a folder and unzip it. Start up ArcGIS Pro, then choose *blank templates / map*. Set the folder path for your project and give it a name. Having created your project, in the catalog panel on the right, right click on ‘folders’ and ‘add folder connection’ to connect to the folder where you downloaded the zip file for this exercise. If you have not done so already, unzip the data to this location. You can then drag and drop data from the catalog panel into the main map window to open it up, so it appears in the layers panel on the left.

Prepare the general practice data:

Open up the **surgeries** data file in ArcGIS pro. 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*.



Input Table: surgeries

Output Table: surgeries_Stats

Statistics Field(s)

Field	Statistic Type
QOFCASES	Sum

Case field: ZIPCODE

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**.

Open up the output table and have a look at it. Note that you can navigate between geoprocessing and catalog panels on the right via tabs at the foot of the screen. You may need to navigate to the folder containing output **practices** table and drag and drop it onto the map window if it does not display automatically. When you open it up, you should see records showing a count of the number of input rows sharing the same postcode, a summed count of heart disease cases, plus the **zipcode /** postcode.

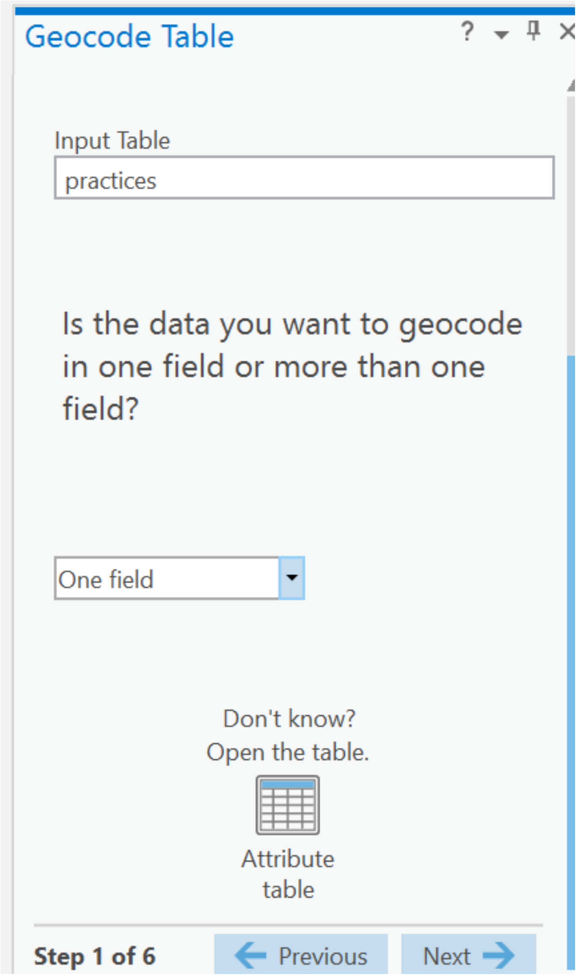
Geocoding post codes using a ready-made, online geocoding service

ArcGIS Pro allows for three types of geocoding:

- one type where place-names to be geocoded are uploaded into the cloud for geocoding against a reference data set held in the cloud. We will explore this option first in this practical.
- a second type where a reference data set is purchased under licence from ESRI, then used in a secure setting to geocode records without compromising data protection (because records are not being passed to a remote server for processing). We will not explore this option here.
- a third where the user creates their own geocoding tool using their own reference data set. We will do this here too, so as get a better understanding of how the process works.

Let us start with a ready-made, online, geocoding service – ESRI's ArcGIS World Geocoding Service. You should hopefully already be logged in to ArcGIS Online, but if not, log in now via the *sign in* link at the top right of the ArcGIS Pro screen. Geocoding takes up credits, which you would normally pay for, though we have a university deal with ESRI. The pricing model is here:
https://pro.arcgis.com/en/pro-app/latest/tool-reference/appendices/geoprocessing-tools-that-use-credits.htm

Next, if you right-click on the **practices** table in the left-hand panel, you should be able to choose *geocode table*, which will open up a guided workflow that you can follow via the buttons at the bottom. Key steps are as follows:



Geocode Table

Input Table: practices

Is the data you want to geocode in one field or more than one field?

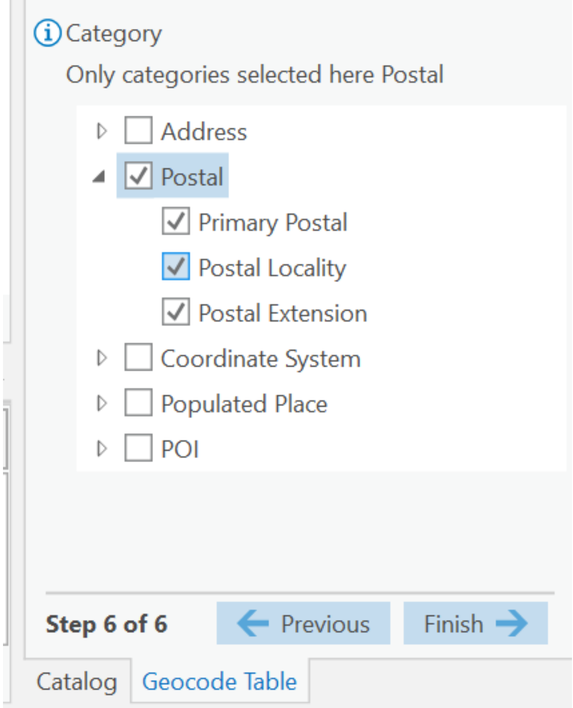
One field

Don't know? Open the table.

Step 1 of 6

2. Choose your locator, a locator in crude terms being one or more reference data sets and a related protocol for carrying out geocoding. Provided you have logged into ArcGIS Online, you should be able to choose *ArcGIS World Geocoding Service* here (if you are not logged in, as this online service requires credits, you will not be able to use it).
3. We will be geocoding the practices table using one field.
4. Set the field for geocoding to be **zipcode.**

4. Choose a location for the output layer that will be produced, e.g. **geocoded_practices.** You can set the *preferred location type* to be either **routing location** or **address location**. The latter is the location of the front door mailbox for the most centrally located property in the unit postcode; the former is the point on the road network closest to this location.
5. Set the country to be the UK. This will speed up geocoding, restricting the reference data set to the subset of spatial features in the UK.
6. In step 6, you can further restrict the reference data set to exclude anything other than postal geography locations (e.g. poi, or 'points of interest' such as attractions and hotels, can be excluded):

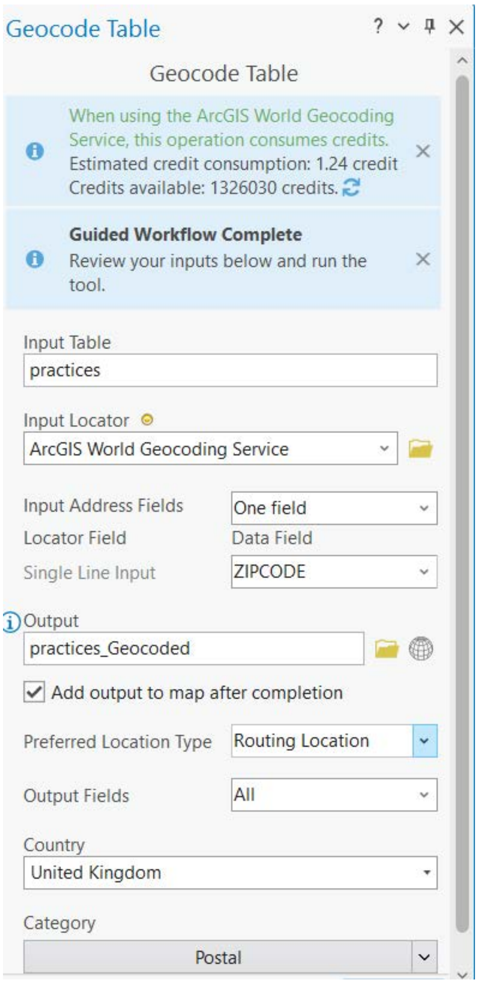


Category

Only categories selected here Postal

- □ Address
- ☑ Postal
  - ☑ Primary Postal
  - ☑ Postal Locality
  - ☑ Postal Extension
- □ Coordinate System
- □ Populated Place
- □ POI

When you hit finish, you should find that the *geocode table* tool (which you could also run via the geoprocessing panel) has been auto-filled for you. Note that it includes an estimate of the number of credits that this look-up will use.



Geocode Table

Geocode Table

When using the ArcGIS World Geocoding Service, this operation consumes credits.
Estimated credit consumption: 1.24 credit
Credits available: 1326030 credits.

Guided Workflow Complete
Review your inputs below and run the tool.

Input Table: practices

Input Locator: ArcGIS World Geocoding Service

Input Address Fields: One field

Locator Field	Data Field
Single Line Input	ZIPCODE

Output: practices_Geocoded

☑ Add output to map after completion

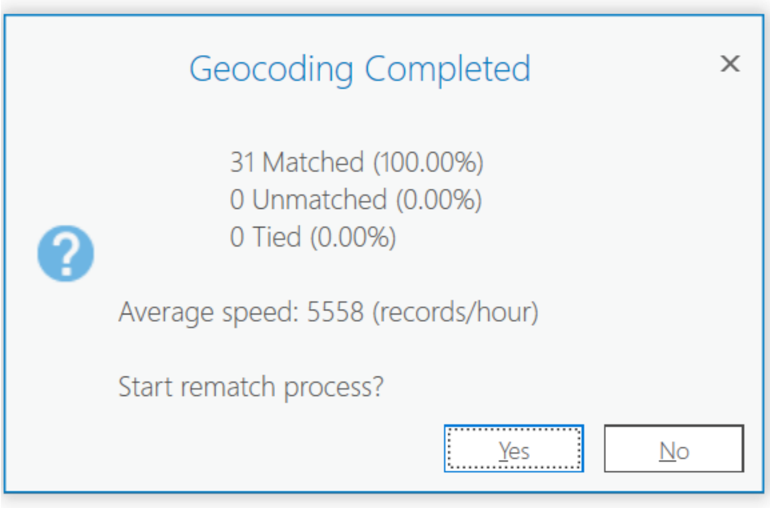
Preferred Location Type: Routing Location

Output Fields: All

Country: United Kingdom

Category: Postal

Now try running the tool. Hopefully, you will see a report something like this:



Geocoding Completed

31 Matched (100.00%)
0 Unmatched (0.00%)
0 Tied (0.00%)

Average speed: 5558 (records/hour)

Start rematch process?

It tells you that the 31 unit post codes have all been successfully matched to the reference data set of unit post codes with locations. As you do not have any unmatched post codes, for now you can say ‘**No**’ to *start rematch process* (we will return to this feature later on). You should hopefully now see a point map of our GP practices – if you cannot, navigate to the location where you placed the output, and then drag the layer into the map window.

How well has the geocoding worked? One way we can examine this is to open up the attribute table of the output layer. If you open up the table, you should see a large number of new fields and if you scroll to the right, then the original attribute fields from the input table (such as heart disease cases) appear there. The help here (see https://pro.arcgis.com/en/pro-app/latest/help/data/geocoding/what-is-included-in-the-geocoded-results-.htm) has plenty of guidance on the many fields that have been added. However, there are a few key fields to keep a close eye on:

- **Status**: tells you the outcome of geocoding. '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. A 'T' means that the record was matched to two or more equally similar corresponding records in the reference data set, or 'tied'. All of our records are 'M' – matched.
- The **Match_type** indicates whether the match was made automatically by ArcGIS ('A'), manually through an interactive table ('U'), or manually through an interactive map ('PP' – pick by point).
- **Addr_type**: This tells you about the level of spatial precision of the match using international terms that could be applied in any country. For a postal geography as the reference data set, 'Postal' indicates a medium precision match, whilst 'PostalExt' indicates a higher precision, more spatially detailed match. In a UK context 'Postal' indicates a postcode sector match (e.g. SO17 1), whilst 'PostalExt' indicates a unit postcode match (e.g. SO17 1BJ).
- **Score**: The score indicates (on a 0 to 100 scale) how closely the post code string in the input table matched to the corresponding post code string in the reference data set. 100 indicates a perfect match; 0 no match.
- **Longlabel** and **shortlabel** provide details of the reference data set record that our postcodes were matched to, i.e. what was the postcode 'in the cloud' that matched to the practice one?

Question: Based on this information, can you work out how well our geocoding exercise has worked? [See next page for answer]

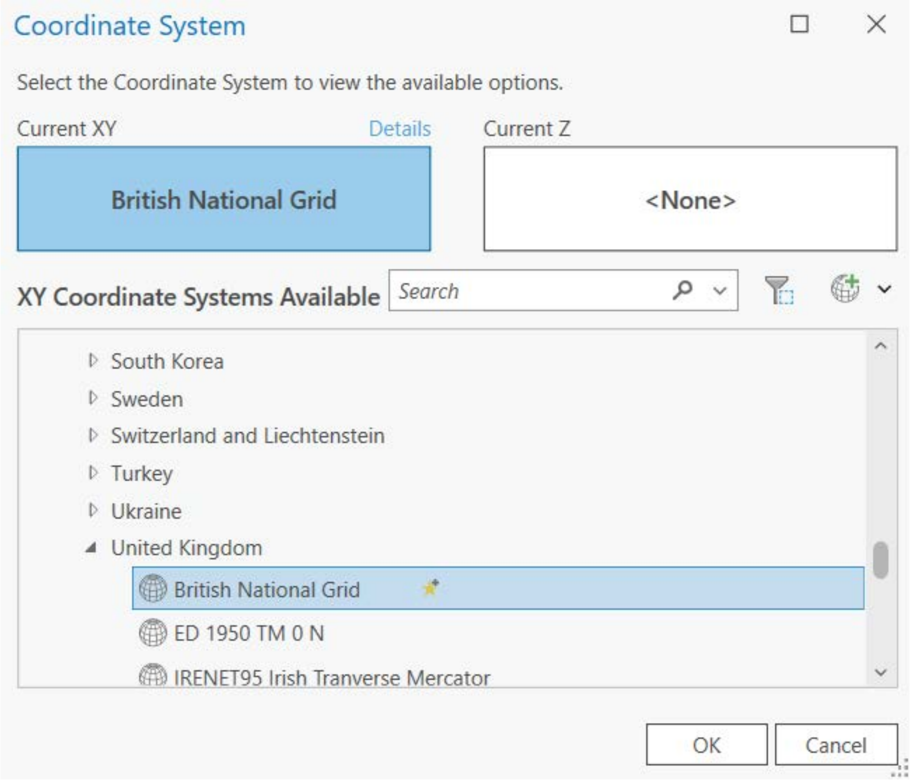
Answer: You should find that there are 4 postcodes with inexact matches to the postcode sector rather than unit postcode level. They have match scores less than 100%. In the absence of a matching unit post code, they have been matched to a postcode sector, the next level up the hierarchy.

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

Having seen a ready-made geocoding service, let us now create one of our own construction.

To do this, first open up the **Codepoint.csv** file within ArcGIS Pro. 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 (note: not the whole of the UK – Northern Ireland has its own coordinate system). Close down the table and map these points as follows:

- Click on the *Map* menu at the top of the screen, then on the ribbon, choose the *add data* button, then *Points from Table*, which will bring up the *xy table to point* tool.
- Click on the button next to *coordinate system* button:



Coordinate System

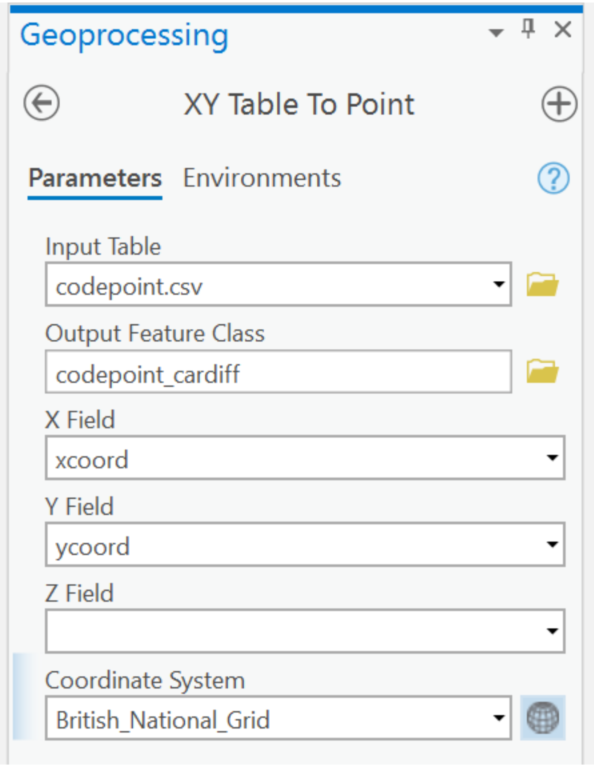
Select the Coordinate System to view the available options.

Current XY: British National Grid

Current Z: <None>

- Search for 'British', and then navigate to *projected coordinate system / national grids / United Kingdom / British National Grid*

- Next, choose the codepoint layer as your input, specify appropriate coordinate fields for mapping the unit postcodes and specify a new output point feature layer:



Input Table: codepoint.csv

Output Feature Class: codepoint_cardiff

X Field: xcoord

Y Field: ycoord

Z Field:

Coordinate System: British_National_Grid

Click on Run 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.

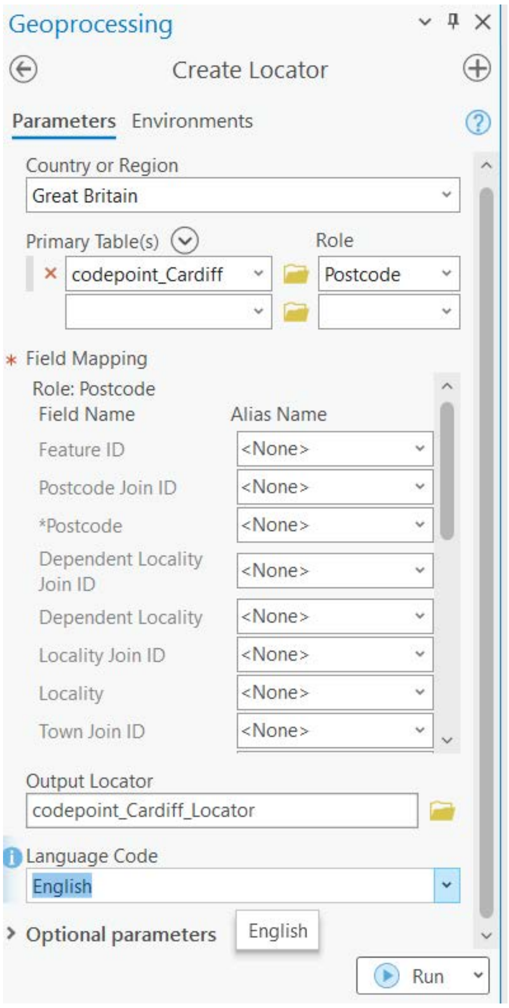
Create an Address Locator

In ArcGIS Pro, 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 geoprocessing panel (via back arrow at top left of dialog box for tool if need be), select Toolboxes (next to Favorites at the top of the geoprocessing panel), then Geocoding Tools, and then choose *Create Locator*. A brief note on some of the other options here before we move on:

- *Create feature locator* supports geocoding where you wish the output of geocoding to be not only points, but polygons or lines.

With *Create Locator*, you can restrict your locator's use to a specific *country or region*:



Country or Region: Great Britain

Primary Table(s): codepoint_Cardiff — Role: Postcode

Field Mapping

Role: Postcode

Field Name	Alias Name
Feature ID	<None>
Postcode Join ID	<None>
*Postcode	<None>
Dependent Locality Join ID	<None>
Dependent Locality	<None>
Locality Join ID	<None>
Locality	<None>
Town Join ID	<None>

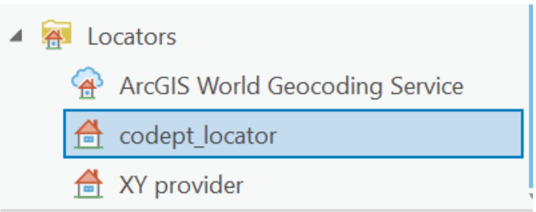
Output Locator: codepoint_Cardiff_Locator

Language Code: English

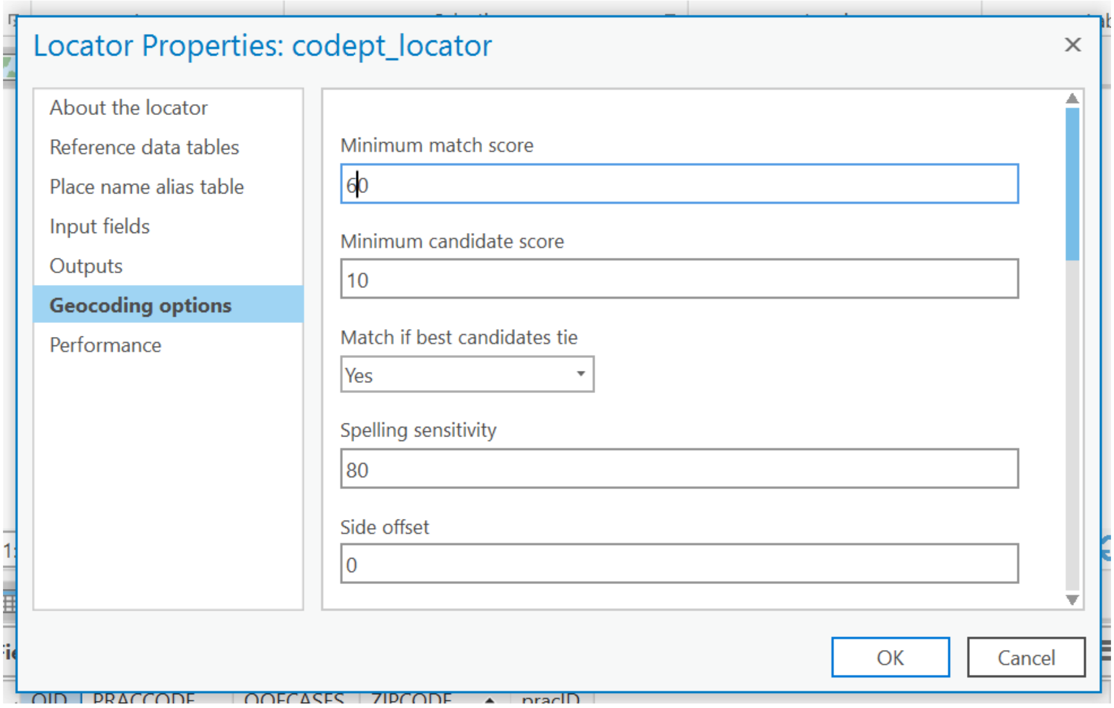
Next, we need to select some *reference data* – in other words, data which will provide us with the point locations of our practices. Select the **codepoint_cardiff** map layer for this. Set its *role* to be **postcode** (note: it is possible to use other ancillary data here – for example where the same street is known by several different names, we can use an additional reference data set as an *alias table* here). You may wish to explore other forms of locator at a later date. 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 **newpostcode.** Finally, at the bottom of the screen, choose a name for your new address locator. Tip: The locator file format is not supported by geodatabases. You must therefore store your locator in a folder, not in a geodatabase. Finally, set the *Language code* to be **English**. Create your locator by pressing Run.

A note about fine-tuning of locators

With some types of locators, such as those for addresses, it is possible to fine-tune the locator after creating it. You can do this via the *locators* area at the bottom of the catalog window, if you right-click on a locator and choose *locator properties:*



This will likely not work for our particular type of locator (a postcode one), as the 'rules' that go with the reference data set can vary and postcode locators do not support such rules. However, with some locator types, it is possible to specify *Minimum match scores*, *minimum candidate scores*, and adjust for *spelling sensitivity*.



Locator Properties: codept_locator

Minimum match score: 60

Minimum candidate score: 10

Match if best candidates tie: Yes

Spelling sensitivity: 80

Side offset: 0

- 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

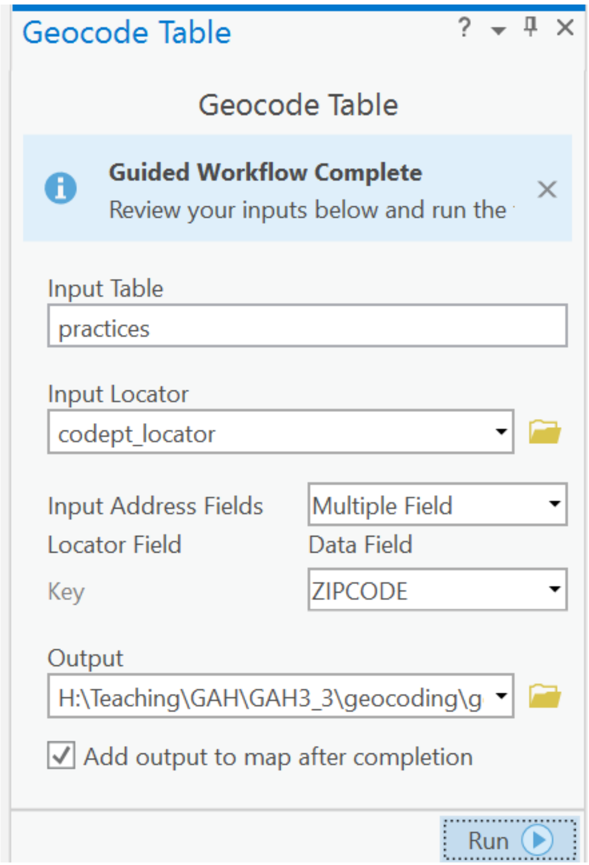
and a **practices** postcode to match them together. If the score for word similarity is lower than this, then ArcGIS will not link our practice data to one of our point locations.

- 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.

All of these metrics relate to string distance, the degree of similarity between text strings. The exact way in which ESRI measures this is not in open code, but as an example, Levenshtein distance is a string distance metric. To compare two strings, Levenshtein distance is the number of text edits (insertions, deletions, replacements) needed to make the strings identical.

Geocode the practice postcodes:

We are now in a position to geocode our **practices** again, but this time using our own home-made locator. Right-click on the layer again in the left-hand panel and choose *geocode table*. This should open a guided workflow as before, but this time, in working through the workflow, choose the locator that you just created, rather than ESRI’s one. When prompted, as the *key*, choose **zipcode**. When you are done, hopefully you should see something like this:



Geocode Table

Geocode Table

Guided Workflow Complete
Review your inputs below and run the

Input Table
practices

Input Locator
codept_locator

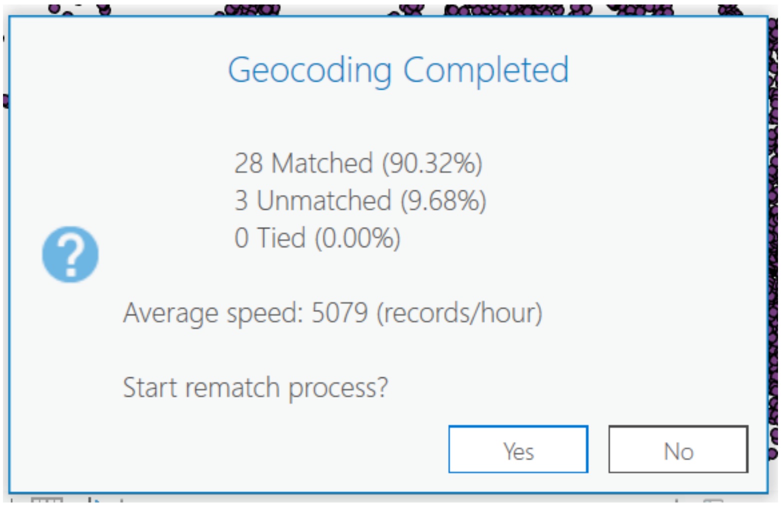
Input Address Fields: Multiple Field

Locator Field	Data Field
Key	ZIPCODE

Output
H:\Teaching\GAH\GAH3_3\geocoding\g

☑ Add output to map after completion

Try running the geocoding tool. When you are done, you should see that our home-made locator has done quite well, though not as well as the ESRI one, matching up 28 post codes:



Geocoding Completed

28 Matched (90.32%)
3 Unmatched (9.68%)
0 Tied (0.00%)

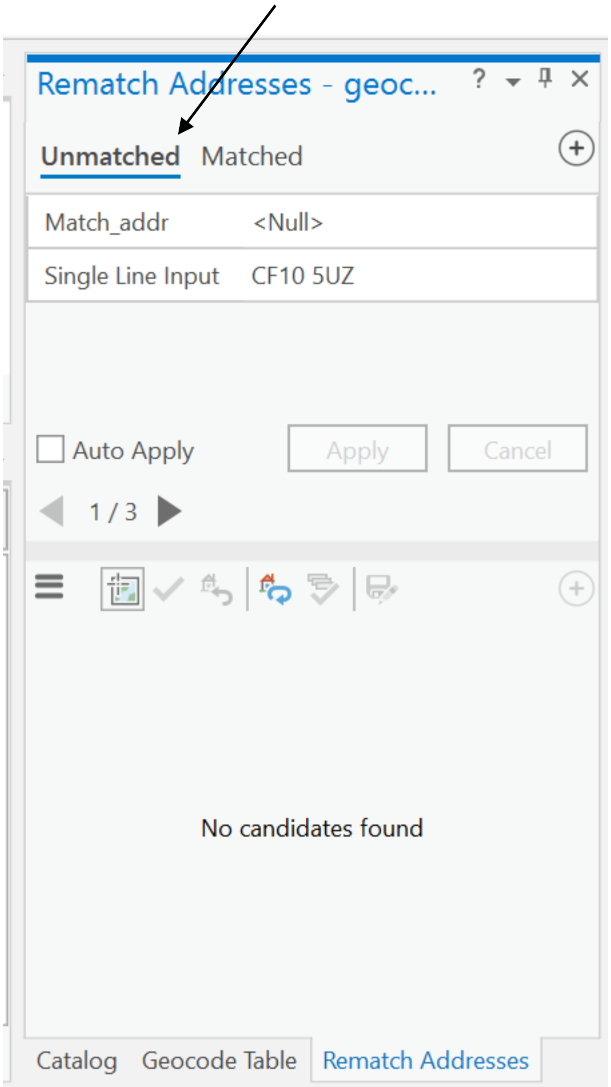
Average speed: 5079 (records/hour)

Start rematch process?

This time, choose YES to start the rematch process, which triggers interactive geocoding.

Interactive geocoding

You will now be presented with a screen for manually reviewing and matching the post codes of surgeries. We had 28 matched practice postcodes, 3 unmatched ones, and none that were tied. The screen breaks these out into different tabs.



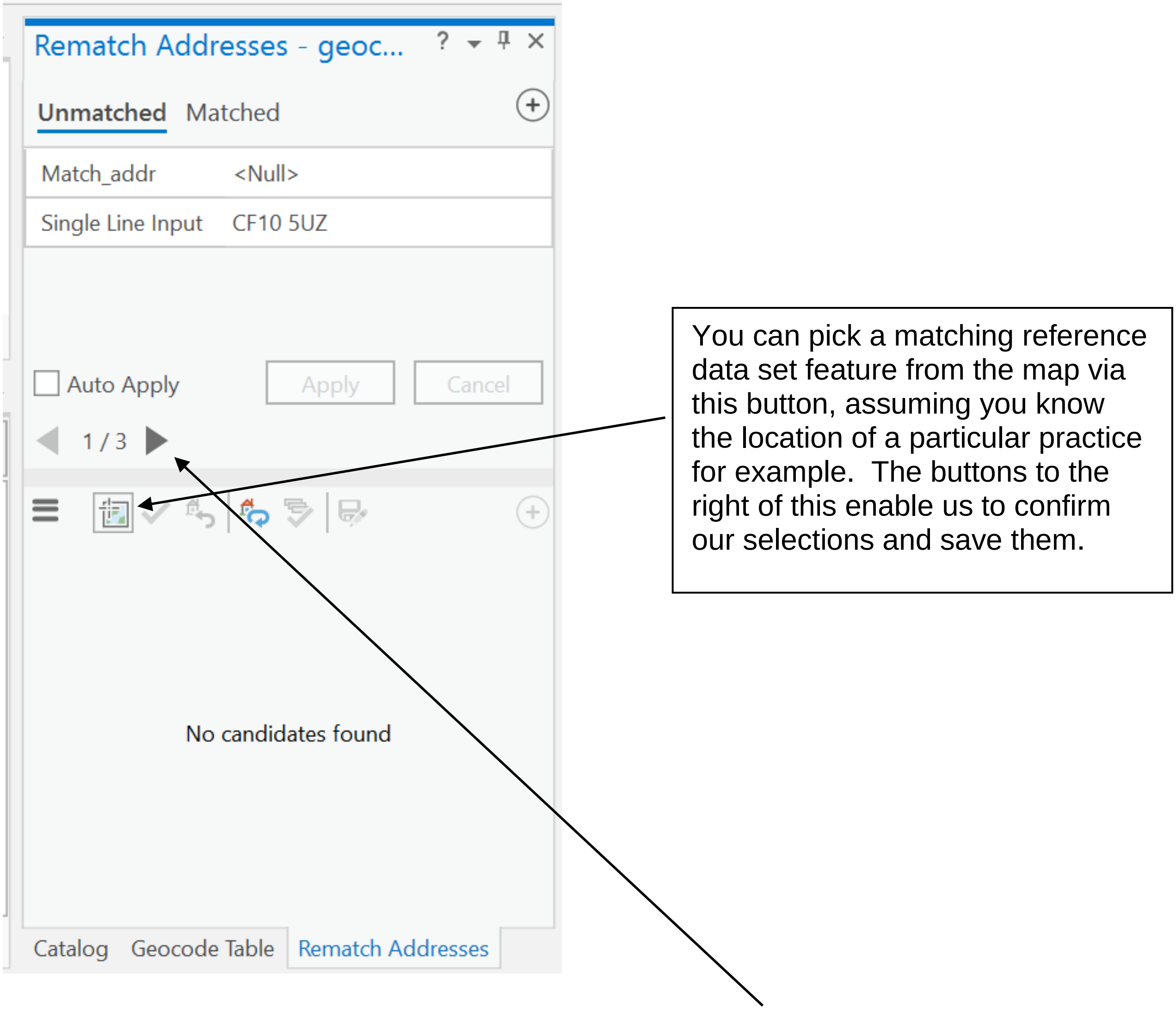
Match_addr	<Null>
Single Line Input	CF10 5UZ

No candidates found

We will focus on the *unmatched* addresses, but if there were incorrectly matched addresses, we could interactively correct these by switching to a review of *matched* addresses. We would see *tied* address matches also listed as a third tab, if there were any, though in our case there are no tied address matches (ties being where two or more reference data set records have equal match scores with the record to be geocoded).

Note that you can open up this screen at any time, even if you say 'no' when finishing off a geocoding operation and being asked about rematching. To do this, right-click on a geocoded feature layer in the left-hand contents panel, choose *data* and then *rematch addresses*. There is also a *rematch addresses* tool, accessible via the geoprocessing panel.

This is how the interactive geocoding panel works:



Match_addr	<Null>
Single Line Input	CF10 5UZ

□ Auto Apply

No candidates found

You can pick a matching reference data set feature from the map via this button, assuming you know the location of a particular practice for example. The buttons to the right of this enable us to confirm our selections and save them.

You can move through the records in each set via these buttons. There are 3 unmatched records for example and you can see their details at the top of the screen and scroll through them.

The bottom part of the screen would normally display ‘candidates’. These are records from the reference data set with post codes that are quite similar to the postcode in the record being geocoded, but not sufficiently similar to be considered a match. Such records have matching scores greater than the minimum candidate score, but lower than the minimum match score. In our little data set, we do not have any candidate records, but if there were, we could review and then select the most promising candidate from the list shown, and match it manually.

In summary, ‘rematch addresses’ gives us a chance to interactively pick reference data set features from a map or reference data set records from a short list, using human intelligence to ‘tidy up’ automatic matches.

The tool is not particular helpful for our particular data set, but it can be very useful in other situations. You can close down the rematch addresses dialog box now.

Reverse geocoding

One final point: You will also find in the geoprocessing panel a *reverse geocode* tool. This does the exact opposite to geocoding, but again using a locator to process data. In other words, it takes a set of locations in a feature layer as an input, then looks up the equivalent addresses, postcodes, or place-names in a reference data set via a locator, adding these fields to the output feature layer. You may wish to experiment with this tool to see what it does too.