Hierarchical Geography with Python

A Very Brief Introduction

Gerald Leung[1]

[1]Public Health Scotland
gerald.leung@phs.scot

March 28, 2022

Background

- We use GP boundaries as an example
  - areas covered are defined with different geographies
  - e.g. postcode districts and sectors
  - ultimately we are only interested in the overall boundary of a GP
- How do we merge geographies at different levels?
  $\Longrightarrow$ `GeoPandas` in Python
- You can find the IPython Notebook here on GitHub

Public Health Scotland

GeoPandas

- An open source project for geospatial data analysis in Python (Jordahl et al., 2020)
- Extends from `Pandas` (Reback et al., 2020), a data analysis package
- Can be used to read and create shape files

Public Health Scotland

NRS Data

We make use of postcode district and sector data from the National Records of Scotland (NRS).

- Shape files
- Contain geometries of districts and sectors
- Read in as `DataFrames` into Python with `Pandas`
- Convert to `GeoDataFrames` with `GeoPandas`

Public Health Scotland

NRS Data

	OBJECTID	District	Shape_Leng	Shape_Area	geometry
0	1	AB10	17466.667741	4.591592e+06	POLYGON ((394256.974 806666.497, 394264.000 80...
1	2	AB11	26086.506732	5.396370e+06	MULTIPOLYGON (((393211.114 805537.072, 393207....
2	3	AB12	73839.045489	7.299857e+07	MULTIPOLYGON (((396494.604 802552.201, 396491....
3	4	AB13	15847.217768	8.770372e+06	POLYGON ((386014.000 803318.000, 385991.953 80...
4	5	AB14	27250.316419	1.754676e+07	POLYGON ((383220.204 804258.703, 383230.799 80...

Figure 1: A segment of `DataFrame` containing district information. Similarly for sector data, with a column representing postcode sectors.

Public Health Scotland

GP Data

- For this example we make use of a few GPs from Lanarkshire (with some modifications):
  - Nalagatla Medical Practice
  - The Craigallian Avenue Practice
  - The Stonelaw Practice
  - Ardoch Medical Practice

Public Health Scotland

Hypothetical GP

- For the purpose of testing, we also create two hypothetical practices, namely Hypothetical One and Hypothetical Two respectively
  - They cover areas defined by a combination of districts and sectors

Public Health Scotland

Data Wrangling

Steps are taken to make sure that the data are presented consistently:

- Separate the comma separated areas of the GPs into individual rows
- Distinguish and separate districts and sectors into different columns

Public Health Scotland

Data Wrangling

Practice Code	Practice Name	Areas
⋮	⋮	⋮
66667	Hypothetical Two	G68,G74 4

Table 1: An example of how the dataset would look like **before** data wrangling.

Public Health Scotland

Data Wrangling

Practice Code	Practice Name	District	Sector
60073	Nalagatla Medical Practice	NaN	G33 6
⋮	⋮	⋮	⋮
60092	The Stonelaw Practice	G76	NaN
66667	Hypothetical Two	NaN	G74 4
66667	Hypothetical Two	G68	NaN

Table 2: An example of how the dataset would look like **after** data wrangling.

Public Health Scotland

Joining the Data

Now that we have our GP (`hyposplit`) and NRS data (`sectors` and `districts`), we merge them together. First merge sector data:

```
# Firstly merge with sector data to get their
# geometries and remove irrelevant rows
merged = pd.merge(hyposplit,sectors,on="Sector",how="
    outer")
merged = merged.head(39)
```

Public Health Scotland

Joining the Data

Then we merge our district data:

```
# Now merge with district data for their geometries
# Again we keep only rows with our GP data
merged = pd.merge(merged,districts,on="District",how="
    outer")
merged = merged.head(39)
```

Public Health Scotland

Convert to `GeoDataFrame`

For a `DataFrame DF`, we can easily convert it to a `GeoDataFrame GDF` with

```
import geopandas as gpd
GDF = gpd.GeoDataFrame(DF, crs="EPSG:4326")
```

where `EPSG:4326` refers to the current coordinate system (latitude and longitude) based on the Earth's centre of mass.

Public Health Scotland

Undissolved Boundaries

- First plot the initial results in `Matplotlib` as a sanity check
- Some boundaries may not be visible due to overlapping
- Notice the boundaries are 'undissolved' - we can still see different levels (sectors and districts) of geographies

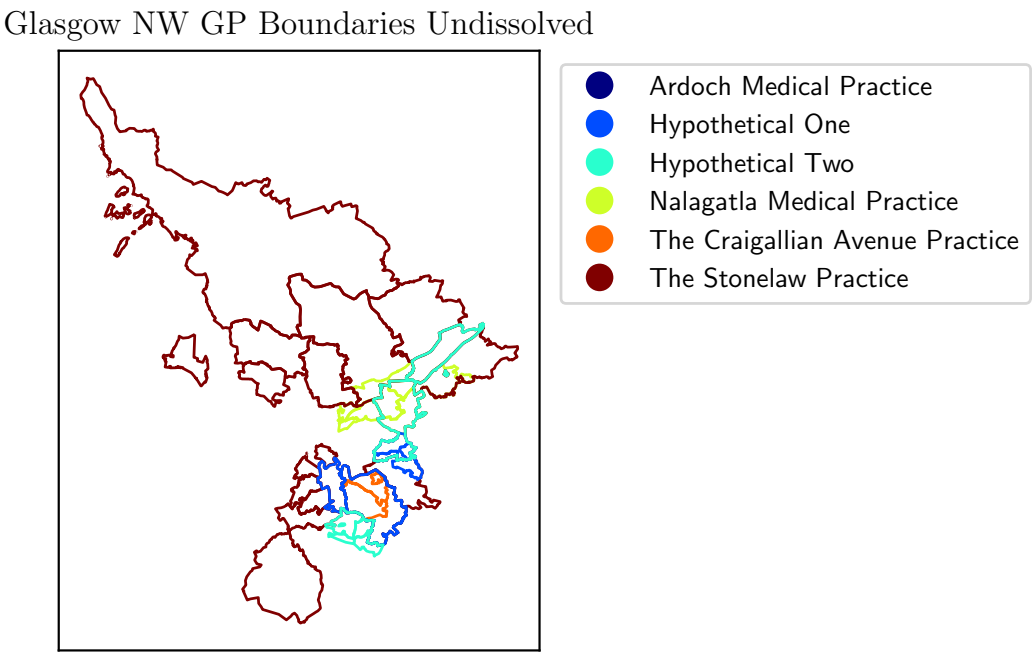

Figure 2: Undissolved Glasgow NW GP boundaries.

Public Health Scotland

Merging Geographies

- We are interested in the overall boundaries of the GPs
- Need to merge the individual postcode districts and sectors of a GP
- We can dissolve the boundaries and merge the geographies

For undissolved `GeoDataFrame uGDF`, we can simply dissolve the geographies by grouping and merging the `Practice Code` column:

```
1 dissolved = uGDF.dissolve(by="
      Practice Code")
```

Public Health Scotland

Dissolved Boundaries

We plot the final results again as a sanity check:

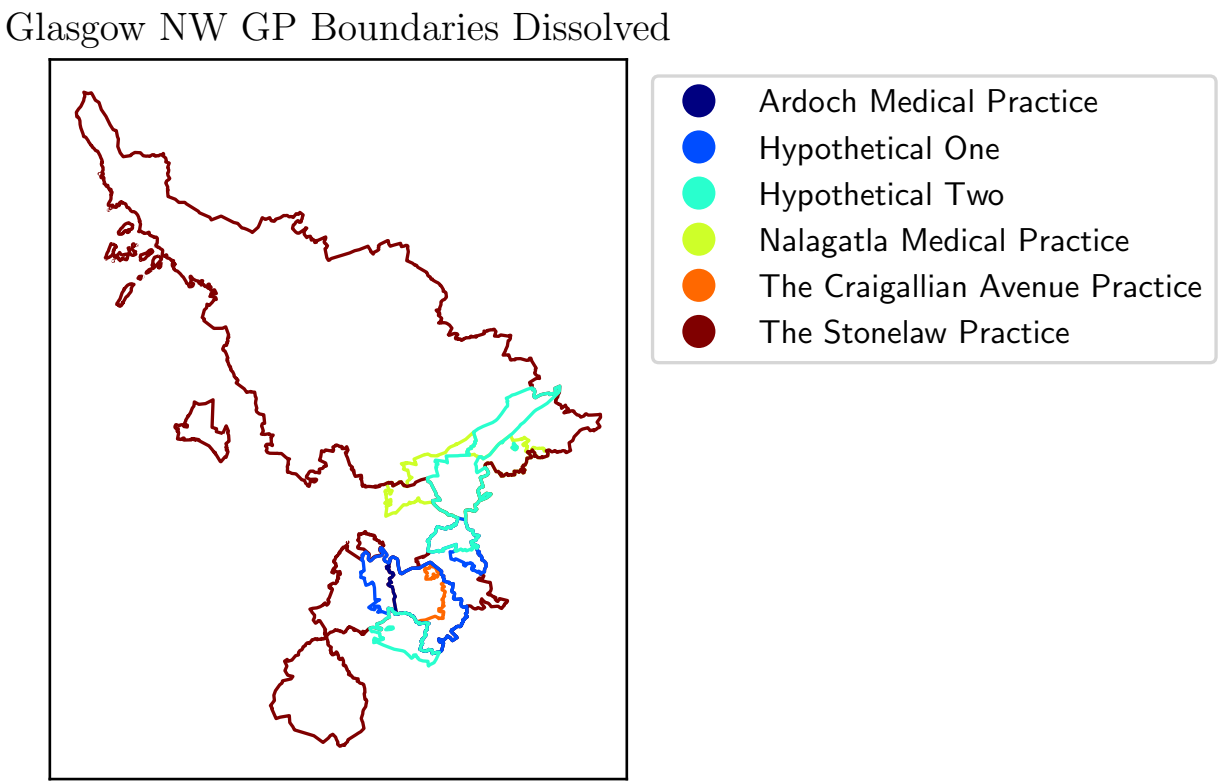

Figure 3: Dissolved Glasgow NW GP boundaries.

Public Health Scotland

Mapping on ArcGIS

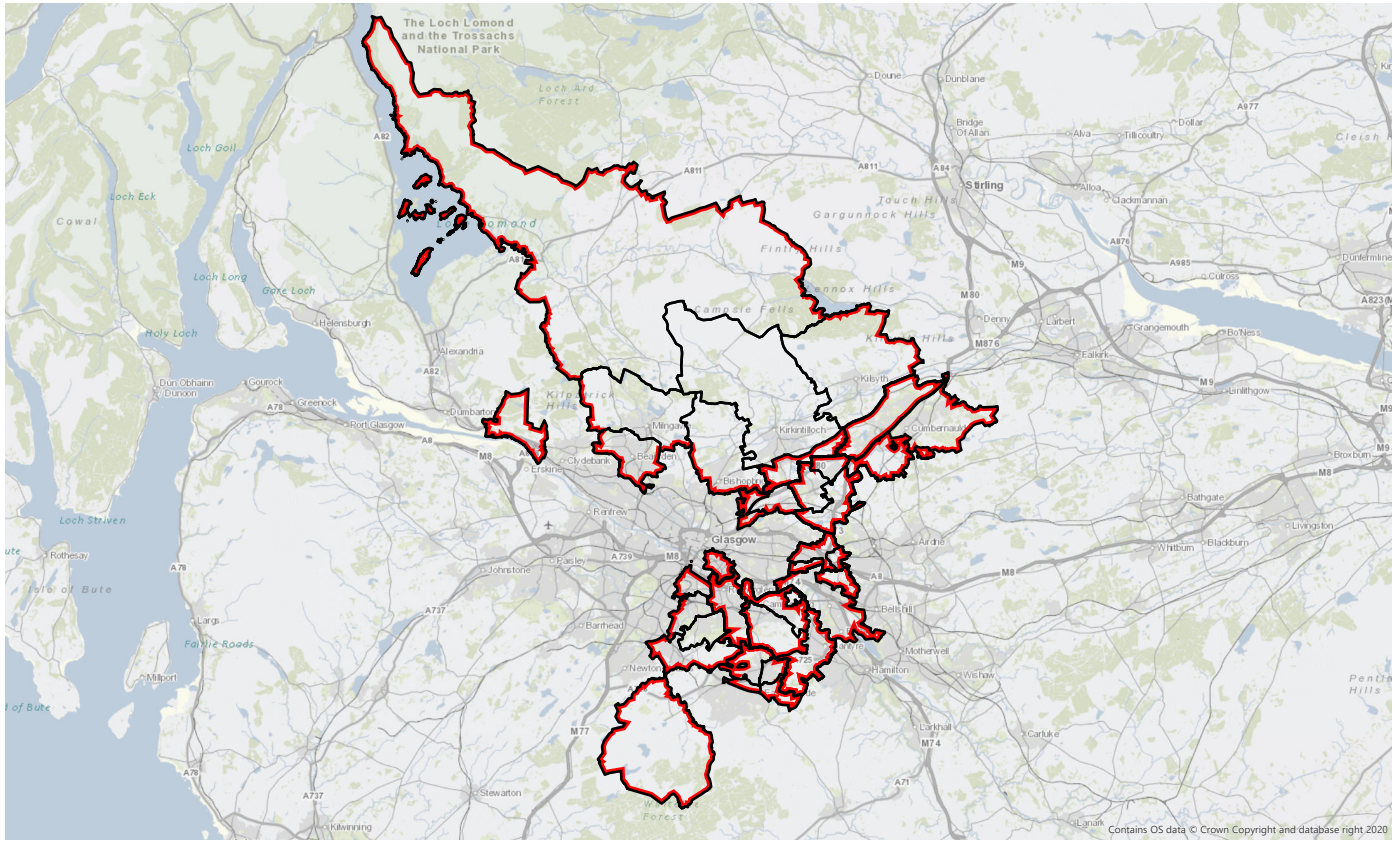

Figure 4: Glasgow NW GP boundaries. Black lines represent all geography levels (sectors and districts). Red lines represent the overall boundaries of a GP (with geographies disolved).

Public Health Scotland

Summary

- `Pandas` and `GeoPandas` in PYTHON provide simple methods to deal with shape files
- Can apply to other scenarios where we have to deal with a hierarchy of geography
- It takes only a few lines of codes in PYTHON for this task so it is relatively simple
- This is however only possible when we are provided with numerical description (e.g. in postcode sectors/districts) of the GP boundaries
  - if we are given a description by words, we will probably have to define the boundaries manually on `ArcGIS`
- Naturally, this task can also be done in `ArcGIS`

Public Health Scotland

References

Jordahl, K., den Bossche, J. V., Fleischmann, M., et al. 2020

Reback, J., McKinney, W., jbrockmendel, et al. 2020

Public Health Scotland