Exploratory data analysis in environmental health
Dr Stéphane Joost, Dr Mayssam Nehme, Noé Fellay

Exercise 7

Geographically weighted regression (GWR) exercise

Overview

In this exercise, you'll explore spatial variability in the relationship between autoantibodies against apolipoprotein A-1 (AAA1) and environmental pollutants. You will use geographically weighted regression (GWR) to uncover how these relationships vary across locations. This assignment will guide you through data exploration, statistical analysis, and spatial visualization using GeoDa, Python, and QGIS. The data comes from the Co-Laus cohort and represents values collected between 2003 and 2006.

Section 1: Data and context

The zip file exercise7.zip is available on Moodle. It contains an ipynb file named "GWR.ipynb" that you will have to complete and it contains a geopackage named CoLausData_ex.gpkg with the following five columns:

- **score2**: Predicts the 10-year risk of developing cardiovascular disease. It considers factors like age, sex, cholesterol levels, blood pressure, and smoking status.
- **aaa1_od**: Autoantibodies against apolipoprotein A-1, which targets HDL cholesterol and is associated with increased cardiovascular risk.
- **cadmium**, **arsenic**, **lead**: Heavy metals collected from urine samples that may contribute to health risks.

Table - CoLausData_ex

	x	y	score_2	aaa1_od	cadmium	arsenic	lead
1	539671.940000	153841.060000	3.500000	0.521400	1.245358	179.940693	7.958225
2	535708.940000	152615.060000	5.500000	0.283050	0.160657	14.255821	1.895741
3	540206.940000	152700.050000	0.600000	0.407850	0.134597	1.713469	0.349462
4	536928.940000	153988.060000	6.600000	0.264700	0.282610	7.140664	1.245299
5	538549.940000	152851.050000	11.300000	0.479600	0.935708	205.635141	3.057090
6	537035.310000	153523.950000	3.400000	0.372750	1.335145	20.190500	4.862590
7	537099.940000	153323.050000	2.600000	0.480300	1.952066	50.023023	2.393082
8	536882.130000	154171.880000	4.500000	0.621550	0.204173	27.469045	0.498092
9	539317.940000	152515.060000	1.800000	0.368950	0.300540	35.074805	1.340350
10	536731.940000	152435.050000	1.000000	1.017300	0.492216	22.956924	2.367277
11	537874.940000	154053.050000	16.299999	0.768750	0.204886	6.453546	6.318133
12	539186.940000	153433.060000	0.400000	1.422800	0.317135	18.680710	0.616336
13	537471.940000	152302.060000	4.800000	0.270200	0.907153	18.698777	6.549011
14	537393.940000	152097.060000	2.200000	0.098850	0.106225	13.707835	0.712859
15	537963.940000	153543.060000	11.300000	0.467400	3.694927	94.070901	4.145863
16	538673.940000	152114.060000	6.400000	0.290000	0.771444	33.752399	2.343154
17	539359.940000	152921.060000	2.700000	0.868150	0.801812	17.886534	0.490570
18	537505.940000	152825.060000	4.400000	2.027400	1.045723	46.185152	1.146989
19	540914.940000	154054.060000	4.400000	0.166300	0.693926	12.998560	1.939414
20	536738.940000	153005.160000	1.700000	0.062550	0.185864	26.929098	0.453532
21	537685.940000	151862.060000	11.200000	1.214400	0.579068	41.261528	2.125727

#row=5161

Exploratory data analysis in environmental health
Dr Stéphane Joost, Dr Mayssam Nehme, Noé Fellay

Section 2: Tasks

1. GeoDa

Task 1: Exploratory data analysis

Use GeoDa to create a scatterplot matrix between score2, aaa1_od, cadmium, arsenic, and lead. Are any variables associated with score2? Are any variables associated with aaa1_od?

Task 2: Subsetting data based on SCORE2

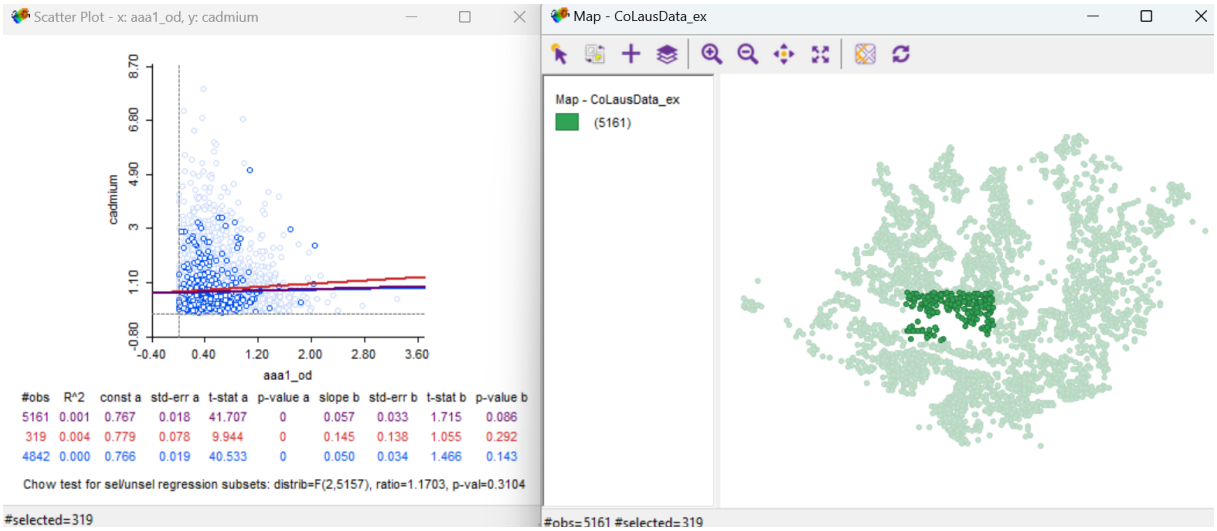
Create a scatter plot between AAA1 (Y-axis) and SCORE2 (X-axis). Using the attribute table, select records where score2 > 7%. Does the relationship between aaa1_od and score2 become significant in this subset? What could be the reasons for this change?

Task 3: Relationship between AAA1 and cadmium

Create a scatter plot with AAA1 on the Y-axis and cadmium on the X-axis. Is the relationship between the two significant and what is the associated p-value?

Task 4: Exploring spatial variability with GeoDa

Open the map with green dots in GeoDa. Use the "Brushing" tool to create a small rectangle (moving window) over different areas.



Notice how the relationship between cadmium and AAA1 changes as you move the window. Can you identify any areas where the relationship between AAA1 and cadmium is significant?

2. Python

Open jupyter notebook in your conda environment

1. **Activate the environment**:
    - Open the Anaconda or Miniconda prompt, then activate the environment you created by running: conda activate gwr_clean_env

Navigate to your working directory:

2. Use the cd command to navigate to the folder where your GWR.ipynb file is located:

   cd C:\XXX...\XXX

Launch Jupyter Lab:

3. In the Anaconda or Miniconda prompt, start Jupyter Lab: jupyter lab
4. Make sure that the environment you are using is “GWR Clean Environment” where we installed the necessary packages.

Now, open the “GWR.ipynb” notebook and complete the following tasks:

Task 5: Understanding GWR outputs

Review the “run_gwr” function (second cell in the “GWR.ipynb” notebook). Explore the parameters: what is the difference between a fixed and adaptive bandwidth? What distinguishes a Gaussian kernel from a bisquare kernel?

What do the p-values, t-values, and beta coefficients represent in the context of GWR? How do they help in understanding spatially varying relationships?

Task 6: Adjusting AAA1 for SCORE2 and running the “run_gwr” function between AAA1 and Cadmium

Specify your file path in the corresponding cell, then adjust AAA1 values for SCORE2 using the cell “Adjusting AAA1 for SCORE2”. Complete the cell where it says ### YOUR CODE Here ###

Why is adjusting aaa1_od for score2 useful in this case?

Run the GWR function by completing the cell “Run the GWR function”. Be careful, the GWR must be computed between AAA1 (dependent variable) and Cadmium (independent variable).

Once done, save the “gwr_results” in the folder of your choice. These results will be used for mapping in QGIS.

3. QGIS

Task 7: Mapping t-values in QGIS

Import the GWR outputs into QGIS. Add OpenStreetMap base layout. Create a graduated map using the t-values with natural breaks classification. You can use the “RdYlGn” color ramp. Which areas seem to have a significant relationship between AAA1 and cadmium?

Task 8: Mapping beta coefficients and creating a map

Using a color ramp of your choice, represent the magnitude and direction (positive or negative) of the beta coefficients for the relationship between cadmium and AAA1.

Overlay areas where p-values indicate statistical significance ($p < 0.05$). You can use a red buffer to highlight these significant areas.

Use QGIS's Print Layout to design a complete map illustrating the effect of cadmium on adjusted AAA1. Include the following map elements:

- **Title, legend, scalebar, north arrow, and source/credits/name**

Examine the spatial distribution of the beta coefficients and the significant areas. Where are the significant beta coefficients located? Discuss what these patterns suggest about the local relationship between cadmium and AAA1.

Task 9: Reflecting on GWR and its limitations

What are the limitations of GWR? How does GWR compare to global regression methods?