Questions and Answers for Lab 3: Crime in the City

In this notebook you will find the questions asked of you in this lab. Beneath each question will be a blank (markdown) cell in which you can write your answer. If you require additional cells use the **Insert** menu to do so, and then use the **Cell** menu to change the **Cell Type** to **Markdown**.

Question 1: How many people must live in a dissemination area before Statistics Canada does not 'suppress' the data? Hint: Review this document and search for “Minimum population.”

• Why does Statistics Canada suppress data?

• What is random rounding and how does that affect the data?

• Would the effect of random rounding on any analysis be greater when working with DA files or with CT files (and why)?

(All of the above combine for 2 Marks)

- 40 people or 250 for income data must live in a dissemination area for it not to be suppressed by Stats Canada.
- Statistics Canada suppress data to protect privacy of individuals living in samller dissemination areas.
- Random rounding is a techique where all figures that are greater than 10 are rounded either up by 5 or down by 5. All figures that are less than 10 are either rounded to 10 or 0. This affects the data because some area with a population of 0 could now be a population of 10 or more. This decreases data precision as by normalizing the data into random multiples of 0, 5, and 10.
- The impact of random rounding on results would be more pronounced in census tracts due to attributes being measured in larger increments and averaged across broader areas.

Question 2:

• What are the min/mean/max median incomes for the DA’s and for the CT’s. Are the differences in the median income values between the DA’s and CT’s as you might expect? Explain. (Hint: One way to explore this is to go to the attribute table, right-mouse click on the field name (`MedHHInc`), and choose an option to visualize the statistics.)

• What is the regression equation that describes the relation between the total number of private households, median household income (/1000), and the number of residential B & E’s for the DA data? The numbers you are looking for can be found in the output for each model (as noted above, these are found via `Analysis` > `History` > `View Details` > `Messages` , which show the coefficients). Are both predictor variables statistically significant (the probability

value, indicated here by a column called `Probability`, should be < 0.05)? In writing your equation only include those variables that are statistically significant at this 0.05 level.

• What is the R^2 value associated with the DA regression equation? Round the value to 2 decimal places.

• What is the regression equation that describes the relation between the total number of private households, the median household income (/1000), and the number of residential B & E's for the CT data? Are both predictor variables statistically significant at the 0.05 level (the probability value in the column associated with each such variable rows should be < 0.05)? In writing your equation only include those variables that are statistically significant at the 0.05 level.

• What is the R^2 value associated with the CT regression equation?

• What do the differing OLS regression results tell us about the effects of resolution (i.e., the scale effect) on the relations between the independent and the dependent variables? (6 Marks)

1. Min for CT is 17,051; Max for CT is 143,155; Mean for CT is 69,472.58; Median for CT is 67,968

- Min for DA is 12,704; Max for DA is 213,504; Mean for DA is 73,048.65; Median for DA is 70,912
- The differences between CT and DA are that min for CT is higher and mean and max for DA is higher. The median is higher for DA as Da contains the smaller area so it is more concentrated on smaller population which tends to be richer than bigger populations.

2. The regression equation that describes the relation between the total number of private households, median household income and the number of residential B & E's for the dissemination area data is: y (B&E) = 0.298519 + 0.004636x (TOT_PRIVATE) + 0.060256z (MEDHHINC1000) + Ɛ (error)
3. The adjusted R-squared value for DA is 0.14.
4. The regression equation that describes the relation between the total number of private households, median household income and the number of residential B & E's for the census tract data is: y (B&E) = 1.725563 + 0.002462x (TOT_PRIVATE) + 0.627737z (MEDHHINC1000) + Ɛ (error)
5. The adjusted R-squared value for CT is 0.15.
6. The differing OLS regression results tells us that the effects of resolution changes with the pattern of the relationship between independent and dependent. This could indicate that the relationship is not consistent and may be influenced by factors related to the scale at which the data is measured. For example, at finer resolutions, certain variables may have stronger correlations with the dependent variable due to more localized effects.

Question 3:

• What are the Moran's I values, the z-scores and the significance levels (p-values) (for I) for the CT and DA regression results?

• Are the residuals spatially autocorrelated? What do the results mean in terms of the distribution of residential B & E's?

• Do you think spatial autocorrelation affects the results of the regression (both the actual regression parameters [e.g., the intercept (a) and the coefficient (b)] and the regression statistics (e.g., the R^2 value)? If so, could you suggest how and/or why? (3 Marks)

	Moran's Index	z-score	p-value
DA	0.275046	28.125382	0.000000
CT	0.315325	6.984460	0.000000

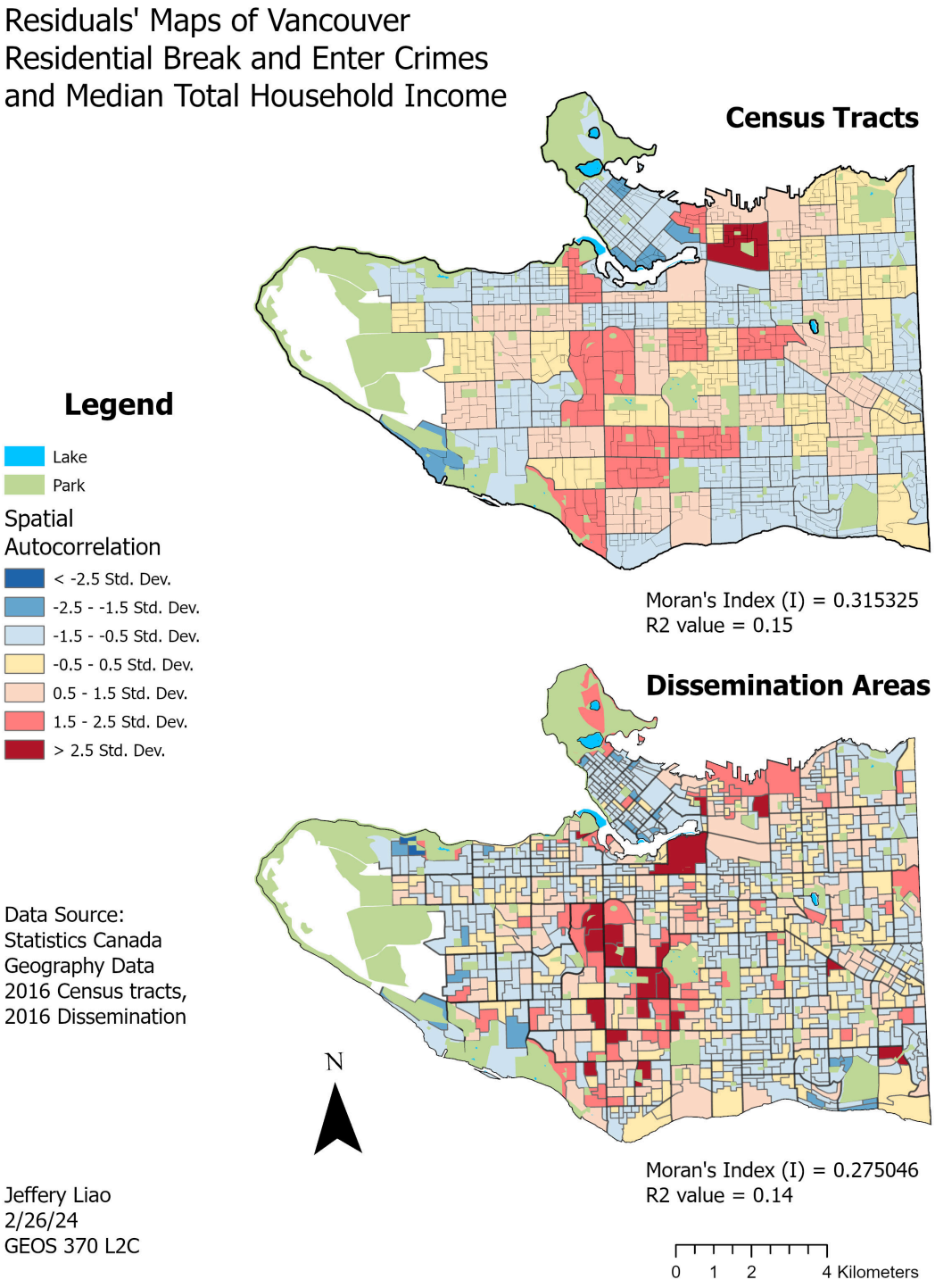
- A Moran's I close to +1 indicates positive spatial auto-correlation (clustering), while a Moran's I close to -1 indicates negative spatial auto-correlation (no clustering). The Moran's Index I values definitely show to be slightly spatially autocorrelated, but certainly not significantly spatially autocorrelated. In terms of residential B & E's, it shows wether the predicted crimes in a neighborhood is greater than, less than, or equal to the observed number of neighborhood crimes. Dark red shows observed crime greater than predicted vaule while dark blue shows observeed crime smaller than predicted value.
- Spatial autocorrelation between two variables affects the result of regression. Regression is the measure of the reationship between the two variables. Measuring regression relies on there being a correlation betweeen the two variables.

Question 4:

• Provide your map.

• Describe the patterns of the residuals—how the two maps present (dis)similar patterns and what the patterns say about the number of residential B & E's relative to the median household income and the total number of households across Vancouver?

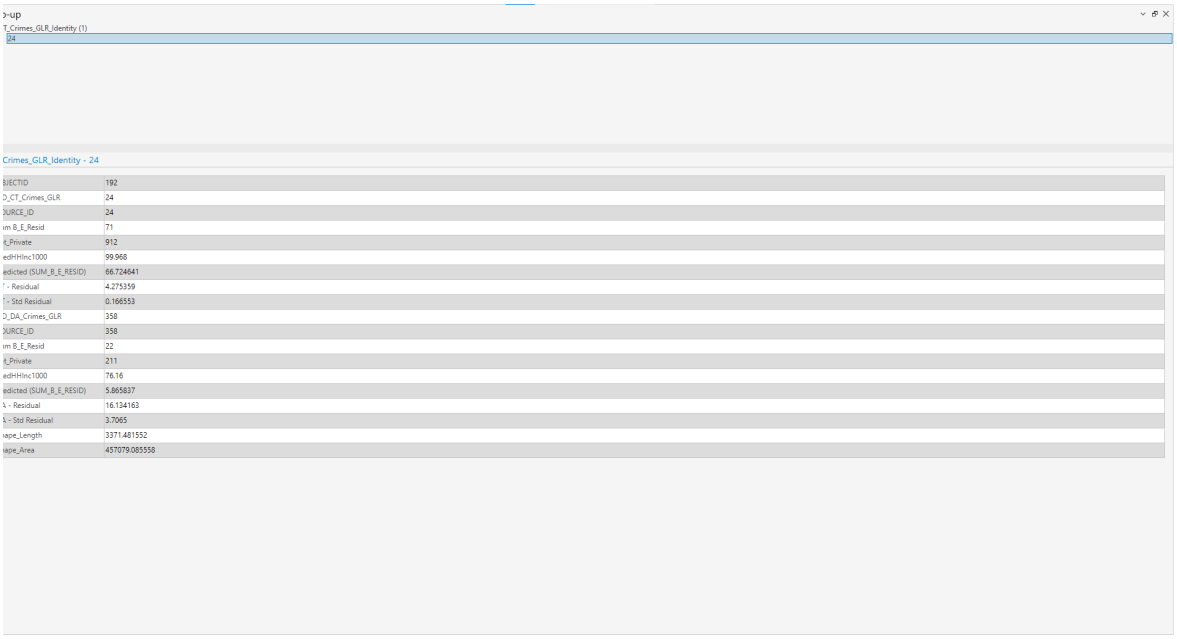
• What impact has scale (i.e., using census tracts versus dissemination areas) had on the results? Visually identify two census tracts that appear to exhibit particularly high variability (that is, areas where the dissemination area map indicates that the single value assigned to the census tract poorly reflects the variability present in the DA's). It might help to zoom into an area and use the `Identify` tool in order to obtain specific values. You could also take a screen-shot of the area and include that in your report. (6 Marks)



- The DA map shows a smaller geograhic unit area compared to the CT map. Overall both maps shows a similar pattern of residuals but CT map shows a gerenal distribution of crime; as class breaks are larger with more similar colors. The patterns show that areas that are darker in color is where crime and median household income are less correlated, while areas of ligher shade are areas where median household income and crime are most correlated. Looking at the DA map there is definitely a more concentrated higher number of residential B&Es relative to the median household income across certain parts of Vancouver. Other dissimilar patterns include looking at downtown Vancouver where there is

homogeneous in the map of CT (median household does not correlate to crime) but in DA map there are pockets where the median household income is strongly correlated with crime.

- The scale affects the map user's understanding of crimes and household income. Using the CT map can lead to users misrepresenting the data and while also masking the true amount of variability between neighborhoods that are close to one another. For example CT 24 has a residual of 4.275359 which shows less correlation but in the DA map the whole area has a residual of 16.134163 which shows high correlation.

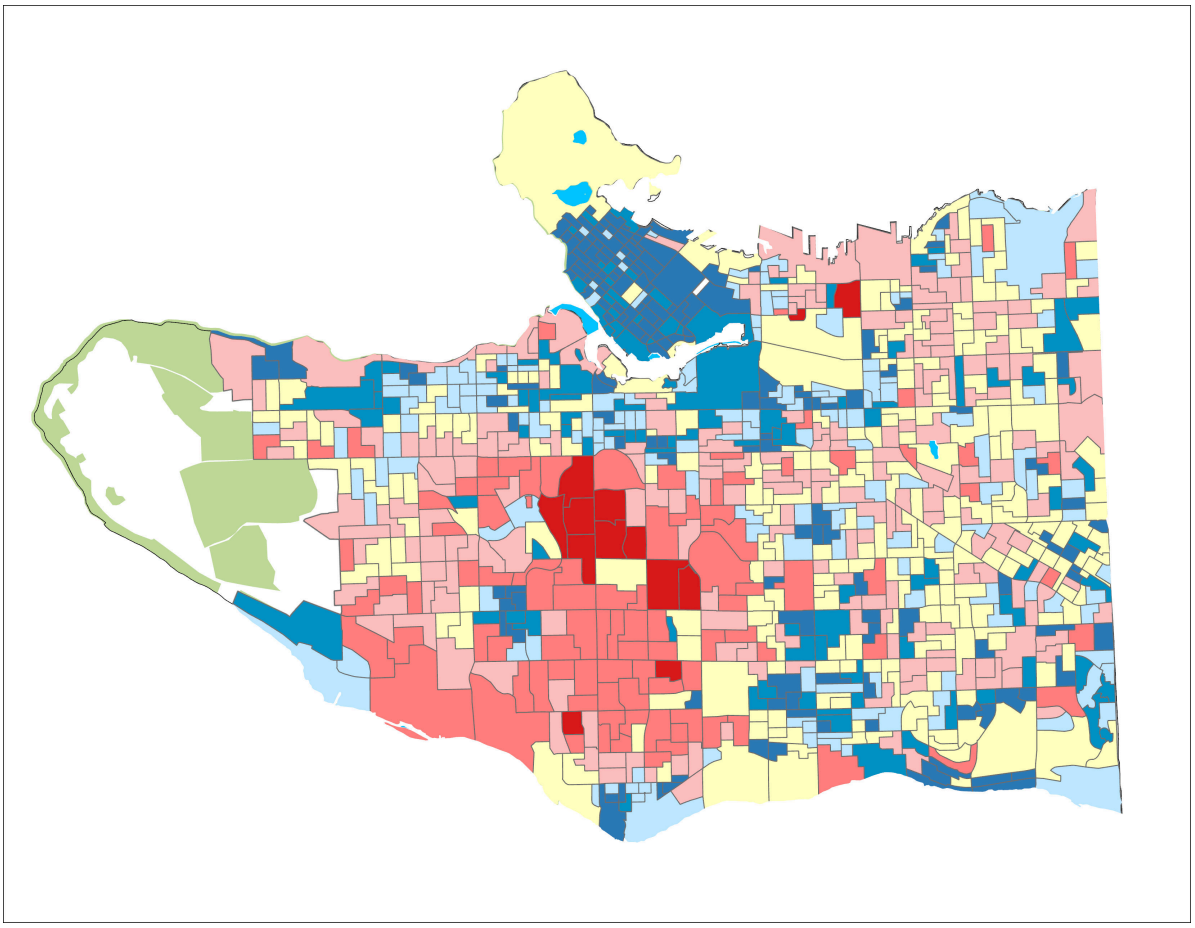


Question 5: What is the Moran's I statistic for the GWR residuals? Is the value still significant? What are the major differences between the two sets of residuals? (1 Mark)

The Moran's I statistic for GWR residual is -0.013565. The value is still significant. The major difference between the two set of residuals is there is a lot more correlation between crime and median household income in the GWR residuals compared to GLR.

Question 6: Does there appear to be some relation between the `LQ` values and the `DA_Crime_GLR` residuals? How could you determine their relation quantitatively? Insert a screen-shot of the `LQ` map in your answer, symbolized using the colour scheme presented above (no need to produce a full formal map layout). (2 Marks)

There seems to be some relationship between the LQ values and the DA GLR residuals. The relation described quantitiatively is that there seems to be equal amounts of spatial autocorrelation between both values. The LQ values has a Moran's I statistic of 0.415008 and the GLR has a value of 0.2.



When done, please submit this notebook as a single PDF to Canvas as before. Note that to convert from 'Marks' (out of 20) to Canvas points (out of 10), we divide by 2.

Congratulations! You've learned a good deal here and the many questions that you no doubt now also have will help guide you as you further explore the complicated, often underappreciated, and yet very enlightening worlds of geographical data analysis!