Introduction to Spatial Data Science - Final Project Guidelines

Luc Anselin

Revised 09/15/2019

The final project should pull together a range of techniques covered in the class to carry out a spatial data exploration. Ideally, this should result in the identification of potential hypotheses or relationships, or lead to discovering the unexpected.

Choose any data set, but make sure you have at least 50 observations. The data can be points, but only if you can analyze the points as discrete units of observation, NOT if the points represent events (such as crimes, accidents, pot holes). In the latter case, you must aggregate the data to an areal unit (such as a community area, census tract, county). You should have at least three variables (more is better).

You do not have to use each and every technique covered in class, but you must include:

- EDA to visualize the non-spatial aspects of the data and explore relationships between variables
- Geo-visualization (mapping) to describe the spatial aspects of the data and to assess potential spatial heterogeneity (or other structure)
- Local spatial autocorrelation analysis to detect clusters (with a sensitivity analysis for the choice of weights, p-values, etc.)
- You may want to reduce multiple explanatory variables (such as census data) to their main principal components (optional)

Your write-up (not including tables and figures) should be 5-10pp. The focus is on the analysis, not on literature reviews, theoretical motivation, etc. **Due December 9, noon**.

Sample Project

Consider the spatial distribution of vehicle thefts in Chicago. You are interested in finding spatial patterns (in the sense of clusters or spatial outliers), the extent to which these remain the same over time, and any potential explanatory factors. You have the individual locations of the points, by time period, and you have a collection of socio-economic data for Chicago community areas (population characteristics, income, housing characteristics) at two points in time.

Your variables of interest are the intensity of vehicle thefts per capita by community area for the two points in time that match the socio-economic data. You are interested in the extent to which any spatial patterns in this variable are the same over time, and if there are potential socio-economic variables that may provide insight into these patterns.

Steps in the Analysis

First, you aggregate the individual points to an areal unit (taking care of any strange data issues), and convert the counts into a per capita rate for each of the time periods. You carry out some initial exploration, using histogram, box plot or other descriptive statistics to assess if there are any strange values (0, or other improbably values). Also, consider some choropleth maps to identify outliers and general patterns (you may want to smooth the rates, if appropriate).

You compare the per capita rates for the two years in a scatter plot (maybe LOWESS curve if there is a suggestion of non-linearity), and you assess the extent to which the spatial patterns coincide with a simple co-location map. You use brushing and linking to identify any possible spatial structural breaks (spatial heterogeneity).

You carry out an initial assessment of potential explanatory variables by means of a scatter plot matrix or a parallel coordinate plot. You may find that too many of the socio-economic variables are highly correlated, so you decide to reduce them to their first two principal components. If you are lucky, these happen to load very nicely along two different dimensions of the data (e.g., one would be population characteristics, the other housing) that are easy to interpret. You include those principal components in the scatter plots and PCP to assess their effect on the theft rates in the two periods.

Next, you move to an identification of local clusters or spatial outliers for the two rates as well as for the principal components. You employ different statistics, different spatial weights and different p-values to identify “interesting” locations that may suggest a potential relationship. You assess the extent to which the clusters overlap or are different. You consider both univariate and bivariate local clusters, and try to interpret the connection between the two by means of linking and brushing other graphs (e.g., a PCP or scatter plot).

If you want to be really creative, you decide you are not happy with community areas as the unit of observation, and instead use census tract data to create your own community areas as spatial clusters based on the principal components of the census data. Then carry out the theft rate analysis for those new regions.

If you cannot find any meaningful patterns, while regrettable, it is nonetheless possible. In such an instance, make sure you explore a full range of techniques before concluding that nothing is there. Reflect this exploration in your discussion.

You can do everything in GeoDa, but you are also free to use R or Python, if you prefer. However, if you don’t use GeoDa, make sure to mimic the type of results you would get in GeoDa and include a description of the code as an appendix.

Things to watch out for

- Not having a shape file (or similar) for areal units
- Not having enough variables (one is not enough)
- Missing values (spatial analysis does not work well with missing values)
- Zero observations (the techniques covered work well for continuous data, not for counts or binary observations)
- Using individual locations (e.g., from survey data) that are not representative of the areal units to which they are aggregated (in a pinch, you may use such data, but be wary of any meaningful interpretation)
- Using spatially extensive variables (counts), convert to spatially intensive instead
- Un-projected data (in decimal degree lat-lon), use great circle distance, or project to a known projection
- Being too liberal with p-values for local autocorrelation tests
- Identifying clusters in maps without cluster statistics
- Discussing patterns that are not justified by the data/analyses
- Not including the graphs and maps that you discuss in the text
- Showing non-significant locations on a local cluster map created with R or Python