Applied Quantitative Methods for the Social Sciences II
MA in Social Sciences, UC3M, Spring 2026

Assignment 8: Spatial Data II

Instructions:

- Deadline: **April 9, before class**
- Submit your work in a separate folder in your GitHub repository
  - You can include only the R file or additional ones (e.g. pdf with results)
- **Always use comments** in your R code – and use them to answer questions
- You are encouraged to work together, but each person must submit their own code
- Plan is to start Part 1 in class and complete Part 2 at home
- I'll upload a solution file to the website after next class

Contents

1 Part 1: In-Class (Detecting Spatial Autocorrelation)

In this lab we continue working with the `world` dataset from the `spData` package, which you already used in Assignment 7. This time we ask a regression question: does GDP per capita predict life expectancy across countries? And if so, are the OLS residuals spatially clustered? We will use Moran's I and Lagrange Multiplier tests to detect spatial dependence, and then fit a Spatial Error Model (SEM) to address it.

Load the required packages and data with:

```
library(sf)
library(spData)
library(spdep)
library(spatialreg)
data(world)
```

Key variables in `world` used in this assignment:

- `name_long` — country name (long form)
- `continent` — continent
- `gdpPercap` — GDP per capita (USD)
- `lifeExp` — life expectancy (years)
- `pop` — population estimate
- `geom` — geometry column (sf)

1.1 Setup and OLS baseline

a) Load the `world` dataset. Remove rows with missing `gdpPercap` or `lifeExp`, and remove Antarctica. How many observations remain? Log-transform `gdpPercap` and store it as a new column `log_gdp`:

```
world = world[!is.na(world$gdpPercap) & !is.na(world$lifeExp), ]
world = world[world$continent != "Antarctica", ]
world$log_gdp = log(world$gdpPercap)
```

In a comment, report the number of remaining observations and explain why we log-transform GDP per capita. Recall the discussion of skewed distributions and log transformations from earlier in the course.

b) Fit an OLS regression of life expectancy (`lifeExp`) on log GDP per capita (`log_gdp`):

```
ols_fit = lm(lifeExp ~ log_gdp, data = world)
summary(ols_fit)
```

In a comment, report the estimated coefficient on `log_gdp`. What does it mean substantively? Is it statistically significant? What is the R^2?

c) Save the OLS residuals and map them:

```
world$ols_resid = residuals(ols_fit)

library(ggplot2)
ggplot(world) +
  geom_sf(aes(fill = ols_resid), color = "white", linewidth = 0.2) +
  scale_fill_gradient2(low = "#2166ac", mid = "white", high = "#d6604d",
                       midpoint = 0, name = "OLS residual") +
  theme_void() +
  labs(title = "OLS residuals: life expectancy ~ log GDP per capita")
ggsave("ols_residuals_map.pdf", width = 10, height = 5)
```

In a comment, describe the geographic pattern. Do you see clusters of positive or negative residuals? Which regions appear to have higher life expectancy than the model predicts, and which appear lower?

1.2 Spatial weights matrix

a) Create a queen contiguity neighborhood and row-standardized spatial weights. Recall from class (and Assignment 7) that queen contiguity defines neighbors as any polygon sharing at least one point:

```
nb = poly2nb(world, queen = TRUE)
listw = nb2listw(nb, style = "W", zero.policy = TRUE)
summary(nb)
```

In a comment, report how many countries have zero neighbors. Explain why some countries have no neighbors in a contiguity-based weights matrix.

b) Test Moran's I on the OLS residuals:

```
moran.test(world$ols_resid, listw = listw, zero.policy = TRUE)
```

In a comment, report the Moran's I statistic and p-value. Is there statistically significant spatial autocorrelation in the residuals? What does this imply for OLS – specifically, what assumption of OLS is being violated?

1.3 Lagrange Multiplier tests

Recall from class that when Moran's I on residuals is significant, we face a choice: should we use the Spatial Error Model (SEM) or the Spatial Lag Model (SLM)? The Lagrange Multiplier (LM) tests help guide this decision. Run all four tests at once:

```
lm_tests = lm.LMtests(ols_fit, listw = listw,
                      test = c("LMerr", "LMlag", "RLMerr", "RLMlag"),
                      zero.policy = TRUE)
```

```
summary(lm_tests)
```

a) Report the test statistics and p-values for `LMerr` and `LMlag`. Are both significant? In a comment, recall from class what each of these tests is checking: LMerr tests for spatial dependence in the *error term* ($\lambda \neq 0$), while LMlag tests for a spatially lagged *dependent variable* ($\rho \neq 0$).
b) Report the robust versions `RLMerr` and `RLMlag`. The robust tests control for the presence of the other type of dependence. Which is more significant? Based on the LM decision rule from class (if both LM tests are significant, compare the robust versions), which model would you choose: SEM or SLM? Write your reasoning in a comment.

1.4 Spatial Error Model (SEM)

Based on the diagnostics above, fit the Spatial Error Model using `errorsarlm()` from `spdep`:

```
sem_fit = errorsarlm(lifeExp ~ log_gdp, data = world,
                     listw = listw, zero.policy = TRUE)
summary(sem_fit)
```

a) Report the estimated coefficient on `log_gdp` from the SEM and compare it to the OLS estimate. Has the coefficient changed? Report the $\hat{\lambda}$ (lambda) parameter and its p-value. Is it statistically significant?
b) In a comment (2–3 sentences), explain what λ represents in the SEM. Recall from class: the SEM says $\mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$, meaning the error at each unit is partly a function of neighbors' errors. If $\lambda > 0$ and significant, what does this tell us about the structure of the unmeasured factors driving life expectancy?
c) Check whether the SEM has removed the spatial autocorrelation from the residuals. Save the SEM residuals and run Moran's test again:

```
world$sem_resid = residuals(sem_fit)
moran.test(world$sem_resid, listw = listw, zero.policy = TRUE)
```

In a comment, compare this result to the Moran's I on the OLS residuals from question 1.2b. Has the spatial autocorrelation been removed or substantially reduced?

1.5 Distance-based weights: an alternative neighborhood

So far we have used queen contiguity to define neighbors: two countries are neighbors if their polygons share at least one point. But this misses island nations entirely and treats all shared-border pairs as equally connected regardless of distance. An alternative is to define neighbors based on geographic proximity: two countries are neighbors if the distance between their centroids is below a threshold.

a) Compute the centroids of every country and build a distance-based neighborhood in which two countries are neighbors if their centroids are within 300 km of each other. Use the following code:

```
coords = st_centroid(st_geometry(world))
nb_dist = dnearneigh(coords, d1 = 0, d2 = 300)
summary(nb_dist)
```

Note on distances and projections. Earlier in the course we said that computing distances requires projecting to a planar CRS (e.g. UTM). That advice applies when you work within a limited area where a single projection is accurate. Here we have a global dataset: no single planar projection preserves distances everywhere on Earth. The function `dnearneigh()` handles this automatically — when it receives an `sf` object with a geographic CRS (WGS84), it computes great-circle distances on the ellipsoid, which are accurate worldwide. The 300 km threshold is therefore interpreted in kilometers without needing to reproject.

In a comment, compare this neighborhood to the queen contiguity one from question 1.2a. How many countries now have zero neighbors? Is this number higher or lower than before? Why might that be?

b) Create row-standardized weights from the distance-based neighborhood and fit a SEM using the same formula (`lifeExp ~ log_gdp`):

```
listw_dist = nb2listw(nb_dist, style = "W", zero.policy = TRUE)

sem_dist = errorsarlm(lifeExp ~ log_gdp, data = world,
                      listw = listw_dist, zero.policy = TRUE)
summary(sem_dist)
```

In a comment, report $\hat{\lambda}$ and its p-value. Compare the `log_gdp` coefficient and $\hat{\lambda}$ from this model to the contiguity-based SEM in question 1.4a. Are the results substantially different? What does this tell you about the sensitivity of spatial models to the definition of the neighborhood?

c) Run Moran's I on the residuals of this distance-based SEM (using `listw_dist`). In a comment, does this model also succeed in removing spatial autocorrelation from the residuals? Compare to your answer in 1.4c.

2 Part 2: Take-Home (Spatial Lag Model and Model Comparison)

2.1 Spatial Lag Model (SLM)

The Spatial Lag Model (also called the Spatial Autoregressive model, or SAR) posits that the outcome itself diffuses across space: $\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Fit it using `lagsarlm()` from `spatialreg`, with the same formula, data, and weights as the SEM.

a) Report the estimated $\hat{\rho}$ (rho) parameter and its p-value, and report the coefficient on `log_gdp`. Is ρ statistically significant?
b) In a comment (2–3 sentences), interpret $\hat{\rho}$. Recall from class that the SLM captures genuine spatial diffusion: the outcome of unit i is partly determined by its neighbors' outcomes. If $\hat{\rho} > 0$, what does this mean about the relationship between a country's life expectancy and its neighbors' life expectancy?
c) In a comment, explain why the coefficient on `log_gdp` in the SLM output is **not** the marginal effect of GDP on life expectancy. Recall from class: solving $\mathbf{y} = \rho\mathbf{W}\mathbf{y}+\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ for $\mathbf{y}$ gives $\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$. What does this equilibrium matrix $(\mathbf{I} - \rho\mathbf{W})^{-1}$ imply for how a change in x_i propagates through the network?

2.2 Direct and Indirect Effects

a) Compute the equilibrium direct and indirect effects using the `impacts()` function, passing the SLM fit and the spatial weights. Use `R = 500` for simulation-based standard errors (and set a seed for reproducibility). In a comment, report the direct effect, the indirect effect, and the total effect of `log_gdp`. How does the direct effect compare to the raw `log_gdp` coefficient from the SLM output and to the OLS coefficient?
b) In a comment (2–3 sentences), explain the substantive meaning of the indirect effect. Recall from class: the indirect effect captures the spillover from unit i's x to all other units' y, after the spatial feedback loop reaches equilibrium. If log GDP per capita in Country A increases by 1 unit, what does the indirect effect say about life expectancy in neighboring countries?
c) The total effect is larger than the direct effect. In a comment, explain whether this is an expected feature of the SLM. Under what conditions would the indirect effect be larger or smaller? (Hint: think about what happens to the spillover term as ρ approaches 0 versus as ρ grows larger.)

2.3 Model Comparison

a) Compare OLS, SEM, and SLM using `AIC()`. Lower AIC indicates better fit, penalized for model complexity. In a comment, report the three AIC values. Which model has the lowest AIC? Does this agree with your LM-test-based model choice in question 1.3b?
b) Write a short summary paragraph as a comment in your R script (5–8 sentences). Include all of the following: (1) whether spatial autocorrelation was present in the OLS

residuals and how strong it was; (2) which spatial model you selected based on the LM tests and why; (3) how the key coefficient estimate on `log_gdp` differs across OLS, SEM, and SLM; (4) what the SLM implies about life expectancy spillovers across borders; (5) one limitation of using queen contiguity weights for country-level data (think about what the matrix misses).

2.4 Extension: Spatial Durbin Model (optional/bonus)

The Spatial Durbin Model (SDM) nests both SEM and SLM by including a spatially lagged dependent variable *and* spatially lagged covariates. It is estimated with `lagsarlm()` by adding the `Durbin = TRUE` argument.

a) Fit the SDM. Inspect the output: you will see a coefficient for `log_gdp` and a separate coefficient for `lag.log_gdp` (the spatially lagged version, $\mathbf{W} \times \text{log\_gdp}$). Is `lag.log_gdp` statistically significant? In a comment, explain what a significant `lag.log_gdp` coefficient would mean substantively: does it suggest that a *neighbor's* GDP predicts a country's life expectancy, beyond what the country's own GDP already explains?
b) Compare the AIC of the SDM to those of the SEM and SLM from question 2.3a. In a comment, is the added complexity of the SDM (one extra parameter) justified by the improvement in fit? Use the AIC values to support your answer.

3 Data Sources

- World polygons and socioeconomic data: `world` dataset in the `spData` R package (`library(spData)` `data(world)`)

4 Submission

Commit your file to your GitHub repository before the deadline. Put it in a separate folder, e.g. `assignment8`. Make sure your repository is public so I can access it.

Your R script should:

- Be well-organized with clear section headers (using comments)
- Include all code needed to reproduce your analysis
- Include your answers and interpretations as comments
- Save any plots to files (using `pdf()`/`dev.off()` or `ggsave()`)
- Run without errors from top to bottom