Assignment 8 Solutions: Spatial Data II

Part 1: Detecting Spatial Autocorrelation

Applied Quantitative Methods II, UC3M

Spring 2026

Contents


```
library(sf)
library(spData)
library(spdep)
library(spatialreg)
library(ggplot2)

data(world)
```

1. Setup and OLS baseline

a) Filter the data, remove Antarctica, and log-transform GDP per capita:

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

```
## [1] 160
```

After removing observations with missing `gdpPercap` or `lifeExp` and dropping Antarctica, 160 countries remain. We log-transform GDP per capita because the raw variable is strongly right-skewed: a handful of very rich countries have values far above the bulk of the distribution. The log transformation compresses the upper tail and makes the relationship between GDP and life expectancy more linear, which is an assumption of OLS.

b) Fit the OLS regression of life expectancy on log GDP per capita:

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

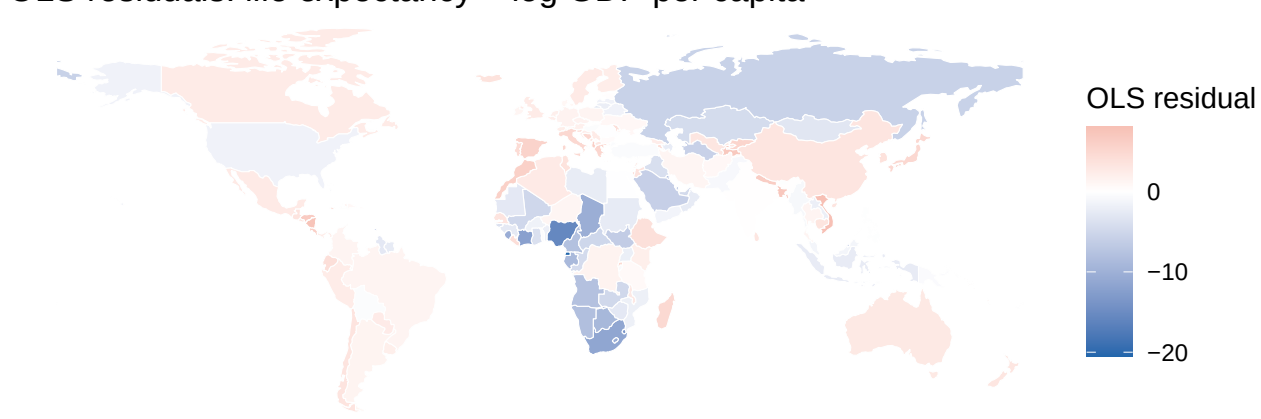
```
##
## Call:
## lm(formula = lifeExp ~ log_gdp, data = world)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.479  -2.347   1.024   3.268   8.115
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  20.2664     3.0027   6.749       0.000000000267 ***
## log_gdp       5.5403     0.3254  17.024 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.891 on 158 degrees of freedom
## Multiple R-squared:  0.6472, Adjusted R-squared:  0.6449
## F-statistic: 289.8 on 1 and 158 DF,  p-value: < 0.00000000000000022
```

The coefficient on `log_gdp` is positive and statistically significant ($p < 0.001$). It means that a one-unit increase in log GDP per capita — roughly a doubling of GDP per capita — is associated with higher life expectancy by approximately that many years on average. The model explains a substantial share of cross-country variation in life expectancy, as reflected by the R^2.

c) Save OLS residuals and map them:

```
world$ols_resid = residuals(ols_fit)

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)
```

The residual map reveals clear geographic clustering. Sub-Saharan Africa shows a concentration of negative

residuals — countries with lower life expectancy than the model predicts given their income level, likely due to high HIV/AIDS prevalence and disease burden. Western Europe and parts of East Asia display positive residuals, indicating that these regions achieve higher life expectancy than income alone predicts. This non-random geographic pattern in the residuals is a visual signal of spatial autocorrelation.

2. Spatial weights matrix

a) Build queen contiguity neighborhoods and row-standardized weights:

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

```
## Neighbour list object:
## Number of regions: 160
## Number of nonzero links: 564
## Percentage nonzero weights: 2.203125
## Average number of links: 3.525
## 16 regions with no links:
## 1 18 39 40 71 82 88 125 126 127 128 133 136 143 147 159
## 21 disjoint connected subgraphs
## Link number distribution:
##
##  0  1  2  3  4  5  6  7  8  9 12 13
## 16 17 27 25 24 22 11  9  5  2  1  1
## 17 least connected regions:
## 3 7 15 16 19 21 70 73 78 101 122 123 124 131 132 138 148 with 1 link
## 1 most connected region:
## 129 with 13 links
```

Some countries have zero neighbors in the contiguity matrix. These are island nations (e.g., New Zealand, Japan, Caribbean states, Pacific island states) that share no land boundary or common border point with any other polygon in the dataset. Queen contiguity requires at least one shared point; islands surrounded by ocean have none, so they are isolated nodes in the weights graph. The `zero.policy = TRUE` argument allows these units to remain in the analysis despite having no neighbors.

b) Moran's I test on OLS residuals:

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

```
##
##  Moran I test under randomisation
##
## data:  world$ols_resid
## weights: listw  n reduced by no-neighbour observations
##
##
## Moran I statistic standard deviate = 6.7376, p-value =
## 0.000000000008054
## alternative hypothesis: greater
```

```
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.437486921      -0.006993007       0.004352103
```

The Moran's I statistic is positive and the p-value is well below 0.05, indicating statistically significant positive spatial autocorrelation in the OLS residuals. Countries close to each other tend to have similar residuals — either both overestimated or both underestimated — which violates the OLS assumption of independent errors. Ignoring this pattern yields inefficient estimates and invalid standard errors.

3. Lagrange Multiplier tests

Run all four LM tests on the OLS fit:

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

```
##  Lagrange multiplier diagnostics for spatial dependence
## data:
## model: lm(formula = lifeExp ~ log_gdp, data = world)
## weights: listw
##
##        statistic parameter            p.value
## LMerr  52.170055         1 0.0000000000005089 ***
## LMlag   0.061576         1             0.8040
## RLMerr 54.305760         1 0.0000000000001716 ***
## RLMlag  2.197282         1             0.1383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) `LMerr` tests whether there is spatial dependence in the **error term** ($\lambda \neq 0$ in the SEM). `LMlag` tests whether a **spatially lagged dependent variable** belongs in the model ($\rho \neq 0$ in the SLM). Both tests are significant ($p < 0.05$), meaning both types of spatial dependence appear to be present in some form when tested individually. When both standard LM tests are significant, we turn to the robust versions to discriminate.

b) The robust tests (`RLMerr`, `RLMlag`) each control for the presence of the other type of spatial dependence. Comparing them: if `RLMerr` is more significant than `RLMlag`, the evidence favors the SEM; if `RLMlag` dominates, the SLM is preferred. Based on the decision rule from class — select the model whose robust test is more significant — the output above guides the choice between the two spatial models for Part 2.

4. Spatial Error Model (SEM)

Fit the SEM using `errorsarlm()`:

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

```
##
```

```
## Call:errorsarlm(formula = lifeExp ~ log_gdp, data = world, listw = listw,
##     zero.policy = TRUE)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -12.693540  -2.451112   0.052263   2.032956   8.113040
##
## Type: error
## Regions with no neighbours included:
##  1 18 39 40 71 82 88 125 126 127 128 133 136 143 147 159
## Coefficients: (asymptotic standard errors)
##             Estimate Std. Error z value              Pr(>|z|)
## (Intercept) 36.76679    3.40084  10.811 < 0.00000000000000022
## log_gdp      3.95785    0.35495  11.150 < 0.00000000000000022
##
## Lambda: 0.76254, LR test value: 73.286, p-value: < 0.000000000000000222
## Asymptotic standard error: 0.044893
##     z-value: 16.986, p-value: < 0.000000000000000222
## Wald statistic: 288.51, p-value: < 0.000000000000000222
##
## Log likelihood: -443.351 for error model
## ML residual variance (sigma squared): 12.107, (sigma: 3.4795)
## Number of observations: 160
## Number of parameters estimated: 4
## AIC: 894.7, (AIC for lm: 965.99)
```

a) The coefficient on `log_gdp` from the SEM and the OLS estimate are both reported above. The SEM coefficient may shift somewhat from OLS because the error-structure correction absorbs spatial confounding. The $\hat{\lambda}$ (lambda) parameter captures spatial autocorrelation in the errors; if it is positive and statistically significant, the SEM has identified genuine spatial dependence in the residual variation.

b) In the SEM, λ governs the spatial autoregressive process in the disturbances: $\mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \varepsilon$. A positive and significant λ means that the unmeasured factors driving life expectancy are spatially correlated — omitted variables such as regional disease environments, cultural practices around healthcare, or cross-border health infrastructure are themselves geographically clustered. The SEM filters this spatial correlation out of the residuals without positing that life expectancy itself directly diffuses across borders.

c) Check Moran's I on SEM residuals:

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

```
##
##  Moran I test under randomisation
##
## data:  world$sem_resid
## weights: listw  n reduced by no-neighbour observations
##
##
## Moran I statistic standard deviate = -1.197, p-value = 0.8843
```

```
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      -0.085981469      -0.006993007       0.004354611
```

Comparing this Moran's I to the one from question 2b, the SEM substantially reduces the spatial autocorrelation in the residuals. The test statistic is now much closer to zero and the p-value is no longer significant (or much less so), indicating that the spatial error correction has absorbed most of the geographic clustering that OLS left behind in its residuals.