Environmental Statistics

Slides - Part 3

Fedele Greco

Department of Statistical Sciences
University of Bologna

Email: fedele.greco@unibo.it

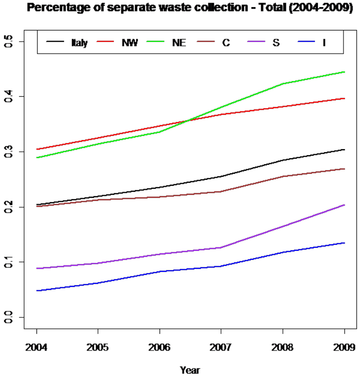
Contents

1 Spatial Regression

Environmental Statistics - Part 3

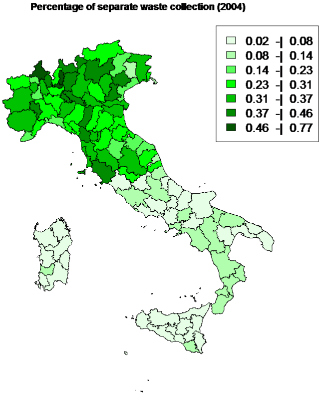
1. Spatial Regression

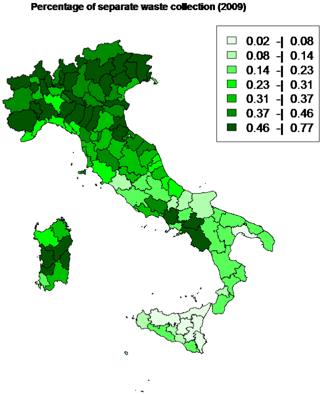
Trend of Separate Waste Collection (SWC) in Italy



An increasing trend is observed from 2004 to 2009. Is this increase uniform? Are there differences in time and space?

Trend of Separate Waste Collection (SWC) in Italy





β-convergence approach

β-convergence approach: “from an economic theoretic point of view it is considered one of the most convincing for exploring the economic convergence of per-capita GDP” (Arbia, 2006).

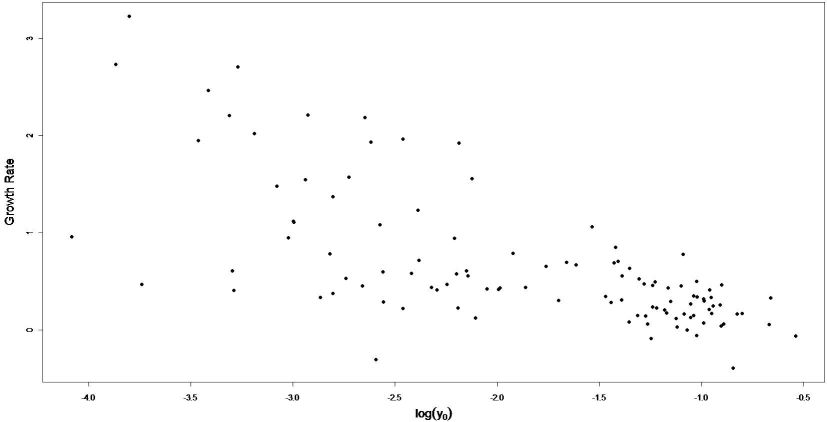
Data: y_{iT} and y_{i0} denote the percentage of SWC at time T (2009) and at time 0 (2004).
The regression model behind the theory of β-convergence is:

$$ln\left(\frac{y_{iT}}{y_{i0}}\right) = \alpha + \beta ln(y_{i0})$$

This is a regression of the growth rate on the starting condition.

Trend of Separate Waste Collection (SWC) in Italy

Growth rate vs log-percentage of SWC in 2004.



Linear regression model

A multiple linear regression model is expressed in matrix form as:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon$$

The regression coefficients are estimated efficiently via OLS:

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

provided that several assumption are met, among these assumptions:

$$\epsilon \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}_n) \tag{1}$$

Moran Test for residual spatial autocorrelation

When performing spatial regression, spatial autocorrelation of the residuals needs to be checked. Indeed, none of the results relating to the estimation and hypothesis testing remain valid if hypothesis (1) is not valid.

If hypothesis (1) is rejected, statistical models suited for managing unexplained spatial correlation are needed. Thus the following hypothesis system needs to be tested:

$$\begin{cases} H_0 : \rho_\epsilon = 0 \\ H_1 : \rho_\epsilon \neq 0 \end{cases}$$

Moran Test for residual spatial autocorrelation

Let $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}$ be the observed residuals.
The test statistic used to test the hypothesis system is:

$$I_\epsilon = \frac{\boldsymbol{e}^T \widetilde{\boldsymbol{W}} \boldsymbol{e}}{\boldsymbol{e}^T \boldsymbol{e}}$$

The moments of the sampling distribution of this test statistic are different from those of the Moran's I we have seen before. The reason is that one needs to take into account that we are dealing with residuals from a model...

Moran Test for residual spatial autocorrelation

Expected value of I_ϵ:

$$E(I_\epsilon|H_0) = \frac{tr(\boldsymbol{M}\widetilde{\boldsymbol{W}})}{n-p}$$

Variance of I_ϵ:

$$V(I_\epsilon|H_0) = \frac{tr(\boldsymbol{M}\widetilde{\boldsymbol{W}}\boldsymbol{M}\widetilde{\boldsymbol{W}}^T) + tr(\boldsymbol{M}\widetilde{\boldsymbol{W}})^2 + \left[tr(\boldsymbol{M}\widetilde{\boldsymbol{W}})\right]^2}{(n-p)(n-p+2)} - [E(I_\epsilon|H_0)]^2$$

where:

$$\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

and *tr* denotes the trace operator.

Linear regression model - R output

```
Call:
lm(formula = log(YT/Y0) ~ log(Y0))

Residuals:
     Min       1Q   Median       3Q      Max 
-1.35347 -0.26120  0.00381  0.20199  1.52184 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.35160    0.11360  -3.095  0.00251 ** 
log(Y0)     -0.54043    0.05503  -9.820  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5067 on 108 degrees of freedom
Multiple R-squared:  0.4717,    Adjusted R-squared:  0.4668 
F-statistic: 96.43 on 1 and 108 DF,  p-value: < 2.2e-16
```

Linear regression model

According to the output:

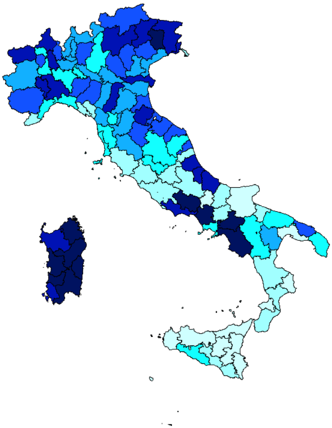
$$\hat{\beta} = -0.54043 \quad se(\hat{\beta}) = 0.05503 \tag{2}$$

Checking for β-convergence requires to test the following hypothesis system:

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$$

On the basis of the output reported in the previous slide, the null hypothesis is rejected, it looks like there is convergence of Italian provinces with respect to their propensity to SWC.
But, are the residuals spatially uncorrelated?

Linear regression model - Residuals



Testing for resituals spatial correlation

$$\begin{cases} H_0 : \rho_\epsilon = 0 \\ H_1 : \rho_\epsilon \neq 0 \end{cases}$$

```
        Global Moran's I for regression residuals

data:
model: lm(formula = log(YT/Y0) ~ log(Y0))
weights: Wlist.til

Moran I statistic standard deviate = 8.5365, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Observed Moran's I        Expectation          Variance
       0.539290100       -0.016485790       0.004238745
```

The null hypothesis is rejected, a spatial regression model is required.

Spatial regression

Ignoring residual spatial autocorrelation causes underestimation of the regression coefficient standard deviation. As a consequence, rejection of the null hypothesis becomes (inappropriately) more likely.

We will discuss two alternatives in the context of spatial regression for areal data:

- the **Spatial Error Model** (SEM)
- the **Spatial Lag Model** (SLM)

Spatial Error Model

The Spatial Error Model is specified as follows:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}$$

$$\boldsymbol{u} = \lambda \widetilde{\boldsymbol{W}} \boldsymbol{u} + \boldsymbol{v}$$

$$\boldsymbol{v} \sim N_n(\boldsymbol{0}, \sigma_v^2 \boldsymbol{I}_n)$$

i.e., a SAR model is used for the errors

Spatial Error Model

From the equation:

$$\boldsymbol{u} = \lambda \widetilde{\boldsymbol{W}} \boldsymbol{u} + \boldsymbol{v}$$

one gets:

$$\boldsymbol{u} - \lambda \widetilde{\boldsymbol{W}} \boldsymbol{u} = \boldsymbol{v}$$

$$(\boldsymbol{I}_n - \lambda \widetilde{\boldsymbol{W}}) \boldsymbol{u} = \boldsymbol{v}$$

$$\boldsymbol{u} = (\boldsymbol{I}_n - \lambda \widetilde{\boldsymbol{W}})^{-1} \boldsymbol{v}$$

Thus the SEM model can be written as:

$$\boxed{\boldsymbol{y} = \boldsymbol{X}\beta + (\boldsymbol{I}_n - \lambda \widetilde{\boldsymbol{W}})^{-1} \boldsymbol{v}}$$

Spatial Error Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma_\epsilon^2((\boldsymbol{I}_n - \lambda\widetilde{\boldsymbol{W}}^T)(\boldsymbol{I}_n - \lambda\widetilde{\boldsymbol{W}}))^{-1}\right)$$

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$$

If the parameter λ was known, model estimation would be easily achieved by means of Generalised Least Squares:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{y}$$

Unfortunately, model estimation is not trivial because λ is unknown. The procedure implemented in R is based on a preliminary estimate of λ, then the GLS estimator is used.

SEM model - R output

```
Call:errorsarlm(formula = log(YT/Y0) ~ log(Y0), listw = Wlist.til)

Residuals:
      Min        1Q    Median        3Q       Max
-0.937289 -0.232150  0.029182  0.193754  1.125681

Type: error
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -0.525216   0.169646 -3.0960  0.001962
log(Y0)     -0.639088   0.071062 -8.9934 < 2.2e-16

Lambda: 0.6515, LR test value: 49.548, p-value: 1.936e-12
Asymptotic standard error: 0.075393
    z-value: 8.6414, p-value: < 2.22e-16
Wald statistic: 74.674, p-value: < 2.22e-16

Log likelihood: -55.51463 for error model
ML residual variance (sigma squared): 0.14077, (sigma: 0.3752)
Number of observations: 110
Number of parameters estimated: 4
AIC: 119.03, (AIC for lm: 166.58)
```

SEM model - R output

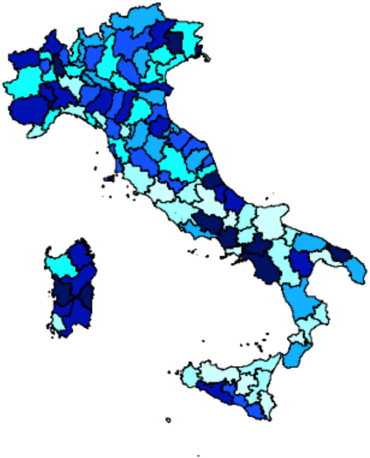
The spatial autoregressive parameter λ is statistically significant, i.e. the hypothesis $H_0 : \lambda = 0$ is rejected.

With respect to the simple linear model, we observe that the standard deviation of $\hat{\beta}$ is increased as expected.

Comparison in terms of the Akaike Information Criterion suggest strong evidence in favor of the SEM model

$$AIC_{SEM} = 119.03 \qquad AIC_{LM} = 166.58$$

SEM model - R output



Spatial Lag Model

The spatial lag model is specified as follows:

$$\boldsymbol{y} = \rho \widetilde{\boldsymbol{W}} \boldsymbol{y} + \boldsymbol{X} \beta + \boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_{\epsilon}^{2} \boldsymbol{I}_n)$$

The SLM is not based on any specific random field model, it consists rather of a technical expedient that seeks to account for spatial dependence between data by adding the spatially lagged dependent variable as a covariate (Arbia, 2006, p.110).

SLM model - R output

```
Call:lagsarlm(formula = log(YT/Y0) ~ log(Y0), listw = Wlist.til)

Residuals:
       Min          1Q      Median          3Q         Max
-1.2792104 -0.2185498 -0.0012651  0.2023904  1.2126790

Type: lag
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.313723   0.100115 -3.1336 0.001727
log(Y0)     -0.362088   0.065727 -5.5090 3.61e-08

Rho: 0.46055, LR test value: 29.238, p-value: 6.4021e-08
Asymptotic standard error: 0.087109
    z-value: 5.287, p-value: 1.2435e-07
Wald statistic: 27.952, p-value: 1.2435e-07

Log likelihood: -65.66963 for lag model
ML residual variance (sigma squared): 0.18229, (sigma: 0.42695)
Number of observations: 110
Number of parameters estimated: 4
AIC: 139.34, (AIC for lm: 166.58)
LM test for residual autocorrelation
test value: 5.6842, p-value: 0.017118
```

SLM model - R output

The spatial autoregressive parameter ρ is statistically significant, i.e. the hypothesis $H_0 : \rho = 0$ is rejected.

With respect to the simple linear model, again, we observe that the standard deviation of $\hat{\beta}$ is increased as expected.

Comparison in terms of the Akaike Information Criterion suggest evidence in favor of the SEM model

$$AIC_{SLM} = 139.34 \qquad AIC_{LM} = 166.58$$

Model selection - Lagrange Multipliers (LM) Test

The Moran's I test for residual spatial correlation does not suggest which model should be preferred when the null hypothesis is rejected.

On the other hand, the LM test can be cast by explicitly expressing the alternative hypothesis either in the form of SEM or SLM.

LM test for the SEM model

The hypothesis system is specified as follows:

$$\begin{cases} H_0 : \boldsymbol{y} = \boldsymbol{X}\beta + \epsilon \\ H_1 : \boldsymbol{y} = \boldsymbol{X}\beta + \lambda\widetilde{\boldsymbol{W}}\boldsymbol{u} + \epsilon \end{cases}$$

An advantage of this test is that one needs to estimate only the model under the null. The test statistic:

$$LM_{SEM} = n^2 \left(\frac{\boldsymbol{e}^T\widetilde{\boldsymbol{W}}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{e}} \right) \frac{1}{tr\left((\widetilde{\boldsymbol{W}} + \widetilde{\boldsymbol{W}}^T)\widetilde{\boldsymbol{W}} \right)}$$

follows a χ^2_1 distribution under the null hypothesis.

LM test for the SLM model

The hypothesis system is specified as follows:

$$\begin{cases} H_0 : \boldsymbol{y} = \boldsymbol{X}\beta + \epsilon \\ H_1 : \boldsymbol{y} = \boldsymbol{X}\beta + \rho\widetilde{\boldsymbol{W}}\boldsymbol{y} + \epsilon \end{cases}$$

The test statistic:

$$LM_{SLM} = n^2 \left(\frac{\boldsymbol{e}^T \widetilde{\boldsymbol{W}} \boldsymbol{y}}{\boldsymbol{e}^T \boldsymbol{e}} \right) \left(tr(\widetilde{\boldsymbol{W}} + \widetilde{\boldsymbol{W}}^T) + \frac{(\widetilde{\boldsymbol{W}}\boldsymbol{X}\beta)^T \boldsymbol{M} \widetilde{\boldsymbol{W}}\boldsymbol{X}\beta)}{\epsilon^T \epsilon} \right)^{-1}$$

follows a χ^2_1 distribution under the null hypothesis.

LM test

```
> lm.LMtests (convergenza.lm, Wlist.til,test="LMerr")         SEM

        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = log(YT/Y0) ~ log(Y0))
weights: Wlist.til

LMErr = 64.2261, df = 1, p-value = 1.11e-15

> lm.LMtests (convergenza.lm, Wlist.til,test="LMlag")         SLM

        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = log(YT/Y0) ~ log(Y0))
weights: Wlist.til

LMlag = 40.0885, df = 1, p-value = 2.427e-10
```

LM test - Robust version

```
> lm.LMtests (convergenza.lm, Wlist.til,test="RLMerr")          SEM

        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = log(YT/Y0) ~ log(Y0))
weights: Wlist.til

RLMerr = 25.1367, df = 1, p-value = 5.341e-07

> lm.LMtests (convergenza.lm, Wlist.til,test="RLMlag")          SLM

        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = log(YT/Y0) ~ log(Y0))
weights: Wlist.til

RLMlag = 0.9991, df = 1, p-value = 0.3175
```