Regression Models for Spatially Autocorrelated Data

Meïli Baragatti

OpenSpat

Table of Contents

1 Sources and Consequences of Spatial Autocorrelation
- Source : interaction
- Source : reaction
- Source : misspecification
- Consequences of the spatial autocorrelation on classical linear models

2 Working example : Las Rosas

3 Spatial Lag Model

4 Spatial Error Model

5 Choosing Between Spatial Lag and Spatial Error models

6 Extended Linear Models

7 Bibliography

Sources and Consequences of Spatial Autocorrelation

When we detect an apparent spatial autocorrelation (on residuals for instance), this spatial autocorrelation may or may not be the result of a spatial autocorrelation.

In 1984, Miron identified three sources of apparent or real spatial autocorrelation :

- interaction
- reaction
- misspecification

Sources of spatial autocorrelation : example

Imagine a population of plants growing in a particular region :

- Y_i measurement of plant productivity (tree height or population density).
- Population is sufficiently dense relative to the spatial scale $\Rightarrow$ productivity measurement may be modeled as varying continuously with the location.
- X_{i1} the amount of light available at location i.
- X_{i2} the amount of available nutrients at location i.

Using these two explanatory variables, the simplest model is :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad \text{with} \quad \epsilon_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2). \tag{1}$$

In matrix notation :

$$\begin{aligned} Y &= X\beta + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma^2 I). \end{aligned} \tag{2}$$

The following three notions can be combined in a same model.

Source : interaction

Spatial autocorrelation induced by interaction occurs when the response variables at different sites interact with each other.

- Negative autocorrelation may occur if trees in close proximity compete with each other for light and nutrients, so that relatively productive tree populations tend to inhibit the growth of other trees.
- Positive autocorrelation would occur if existing trees produced acorns that do not disperse very far, which in turn results in more trees in the vicinity.

If Y is positively autocorrelated, the true underlying model is :

$$\begin{aligned} Y &= X\beta + \rho WY + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma^2 I), \end{aligned} \tag{3}$$

with WY the spatial lag.

Interaction : illustration using simulations (1/2)

We generate a dataset `simu_modlin` satisfying model (2) with $\beta = (0, 0.5, 0.3)$ and a dataset `simu_interaction` satisfying model (3) with $\beta = (0, 0.5, 0.3)$ and $\rho = 0.6$. Each dataset contains 1000 observations and X_1 and X_2 are simulated independently using gaussian distributions.

```
mod <- lm(Ylin ~ X1 + X2)
print(coef(mod), digits = 2)

## (Intercept)          X1          X2
##    -0.00021     0.49979     0.30028

var(mod$res)

## [1] 9.560601e-05
```

Interaction : illustration using simulations (2/2)

```
mod <- lm(Yinter ~ X1 + X2)
print(coef(mod), digits = 2)

## (Intercept)          X1          X2
##       0.028       0.556       0.316

var(mod$res)

## [1] 0.05820449
```

Source : reaction

Spatial autocorrelation induced by reaction occurs when the response variables are reacting to an external factor that varies in space, and when this factor is not taken into account by the model.
For instance if nearby plants are reacting to availability of water (which varies in the 'space').

The inclusion of this external factor in the linear model may be appropriate. It may be sufficient to explain the spatial autocorrelation, and to obtain non-autocorrelated residuals.

For instance, the true model should be :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad \text{with} \quad \epsilon_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \tag{4}$$

with X_{i3} the distance from the river at location i.

Reaction : illustration using simulations (1/4)

We generate a dataset `simu_reaction1` satisfying model (4) with $\beta = (0, 0.5, 0.3, 0.8)$ and X_3 correlated with X_2.

We fit model (4) :

```
print(coef(lm(Yreact1 ~ X1 + X2 + X3)), digits = 2)

## (Intercept)          X1          X2          X3
##      0.0088      0.4837      0.3315      0.7716

mod <- lm(Yreact1 ~ X1 + X2)
print(coef(mod), digits = 2)

## (Intercept)          X1          X2
##        0.51        0.50        1.01
```

X_3 maybe interpreted as a 'spatial' variable, but its role in the model is identical to that of another explanatory variable without any spatial connotation.

Reaction : illustration using simulations (2/4)

We generate a dataset `simu_reaction1` satisfying model (4) with $\beta = (0, 0.5, 0.3, 0.8)$, and X_3 non correlated with X_1 or X_2 but spatially autocorrelated.

We fit model (4) :

```
mod <- lm(Yreact2 ~ X1 + X2 + X3)
print(coef(mod), digits = 2)

## (Intercept)          X1          X2          X3
##       0.027       0.483       0.327       0.768

var(mod$res)

## [1] 1.003906
```

Reaction : illustration using simulations (3/4)

We fit model (1) :

```
mod <- lm(Yreact2 ~ X1 + X2)
print(coef(mod), digits = 2)

## (Intercept)          X1          X2
##       0.046       0.481       0.321

var(mod$res)

## [1] 1.863427
```

The effect of X_3 which is not taken into account in this model is entirely loaded in the error term.

Reaction : illustration using simulations (4/4)

As X_3 was spatially autocorrelated, the result is that the residuals are spatially autocorrelated :

```
lm.morantest(mod,W)


##
##  Global Moran I for regression residuals
##
## data:
## model: lm(formula = Yreact2 ~ X1 + X2)
## weights: W
##
## Moran I statistic standard deviate = 5.6528, p-value = 7.892e-09
## alternative hypothesis: greater
## sample estimates:
## Observed Moran I      Expectation         Variance
##     0.1295740253    -0.0010112445     0.0005336546
```

Source : misspecification

The measured autocorrelation is not due to interaction or reaction but to the incorrect form of the model.
For instance if we assume homoscedastic errors when in fact they are heteroscedastic.

The true model should be (the variance of the errors increases with the amount of available nutrients X_{i2}) :

$$\begin{aligned} Y &= X\beta + \epsilon \\ \epsilon_i &\underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 \times \exp(1 + 2X_{i2})). \end{aligned} \tag{5}$$

In this case, the measured autocorrelation can be induced by the wrong modelisation, it is then an apparent autocorrelation and not a real autocorrelation (this autocorrelation cannot be explained by spatial considerations).

Misspecification : illustration using simulations (1/2)

We generate a dataset `simu_modmiss` satisfying model (5) with $\beta = (0, 0.5, 0.3)$. X_2 spatially autocorrelated and the error variance is an increasing function of X_2.

We fit model (2) :

```
mod <- lm(Ymiss ~ X1 + X2)
print(coef(mod, digits = 2))

## (Intercept)          X1          X2
##    30.95456   -68.36515    84.34902
```

Misspecification : illustration using simulations (2/2)

```
lm.morantest(mod, W)

##
##  Global Moran I for regression residuals
##
## data:
## model: lm(formula = Ymiss ~ X1 + X2)
## weights: W
##
## Moran I statistic standard deviate = 2.3661, p-value = 0.008989
## alternative hypothesis: greater
## sample estimates:
## Observed Moran I      Expectation         Variance
##      0.159988117     -0.014521469      0.005439884
```

The error terms are uncorrelated, but because the error variance is a function of X_2 and high values of X_2 tend to be near other high values of X_2, a test for spatial autocorrelation of the residuals has a high type I error rate.

Consequences of the spatial autocorrelation

Interaction biased estimates of the regression coefficients, the variance of the residuals is inflated $\Rightarrow$ inflated type I or II error rates of certain tests.

reaction If the reaction variable (not included in the model) is correlated to a variable present in the model, the estimate of the coefficient associated with the variable present in the model will be biased.

If the reaction variable (not included in the model) is not correlated to a variable present in the model, but is spatially autocorrelated, the variance of the residuals will be inflated, $\Rightarrow$ inflated type I or II error rates and indication of spatial autocorrelation when none really exists.

Misspecification If the model is misspecified, that can lead to both biased estimates of the regression coefficient and indication of spatial autocorrelation when none really exists.

Table of Contents

1 Sources and Consequences of Spatial Autocorrelation

2 Working example : Las Rosas

3 Spatial Lag Model

4 Spatial Error Model

5 Choosing Between Spatial Lag and Spatial Error models

6 Extended Linear Models

7 Bibliography

Spatial regression models in practice (1/2)

1. Fit the data with a classical linear model like (2).

2. Check the model assumptions on the residuals : normality, homoscedasticity and independence.

   Non-normality histogram, Q-Q plot, Shapiro-Wilk test, Kolmogorov-Smirnov test.

   Heteroscedasticity or the exclusion of a reaction variable plot the residuals against the fitted values, and against the different variables included or not in the model.

   Dependence try to detect a spatial autocorrelation of the residuals : bubble plots, semi-variograms, Moran correlogram, test for spatial autocorrelation of the residuals using the Moran's I.

Spatial regression models in practice (2/2)

3. If we detect some problems on the residuals :

Non-normality the model can be misspecified. Try a transformation of your variable to be explained and/or of your explanatory variables. It can also be the consequence of a relevant explanatory variable forgotten in the model.

Homoscedasticity or the exclusion of a reaction variable take into account this heteroscedasticity in your model.

Dependence check that you have not forgotten a reaction variable, and that you are not in presence of heteroscedasticity. If not, fit a more complicated model with an autocorrelation structure : spatial lag model, spatial error model or an extended linear model with a spatial autocorrelation structure.

Example Las Rosas (1/16)

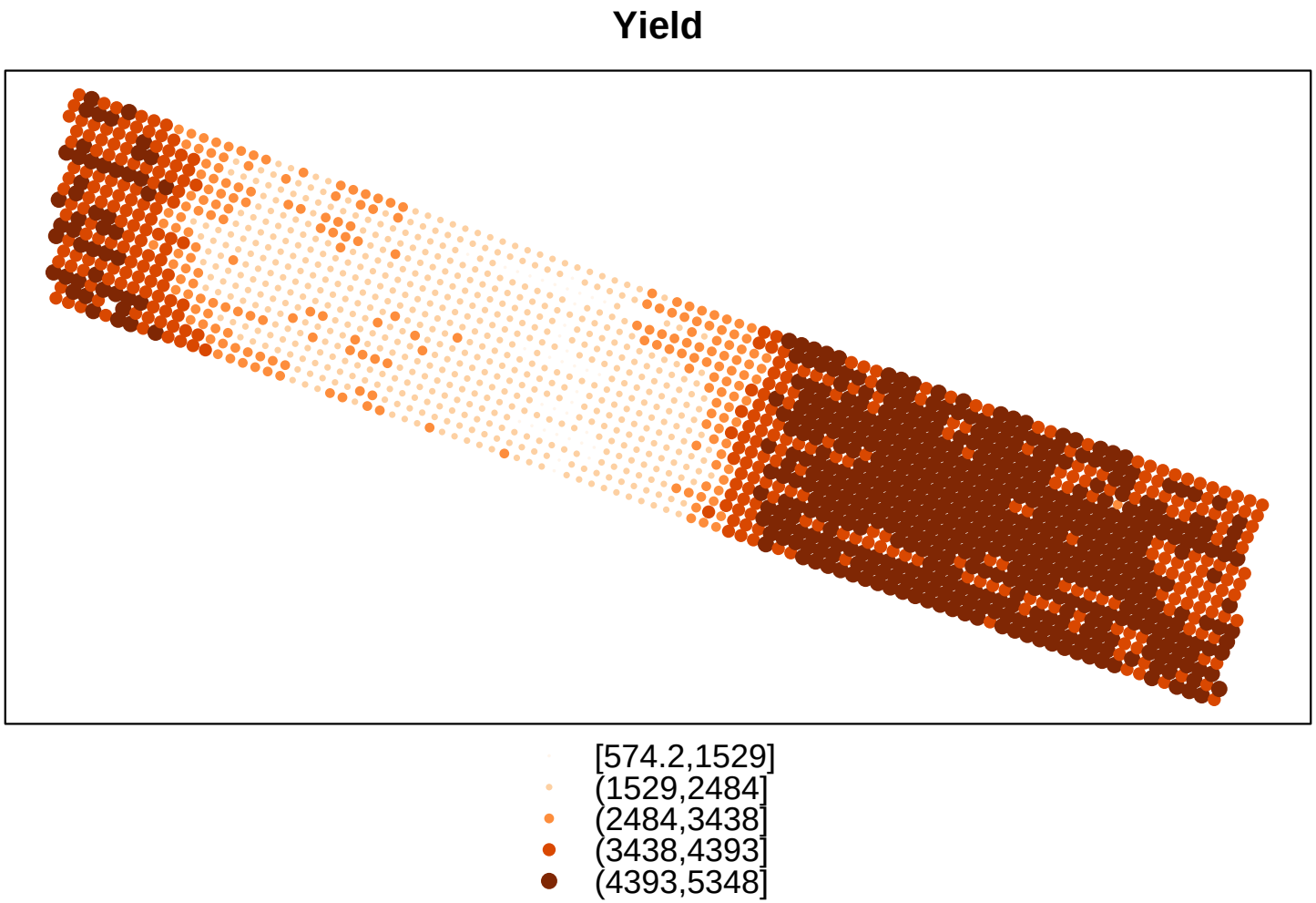
Data set from Anselin et al. 2001.

- Measurements of corn yield over a controlled plot in Argentina. Regular grid approximately 71 cm apart.
- Amount of nitrogen fertilizer that is applied on each location : 6 levels applied along the rows of the field.
- The basic set of information consists of four variables measured at 1704 locations : `YIELD`, `N`, `LATITUDE`, `LONGITUDE`.
- `Xutm` a `SpatialPointsDataFrame` object containing the yield and relevant geographical variables to explain it (`N`, `elev`, `slope`, `slopeX`, `accu`, `aspect` and `hshade`).

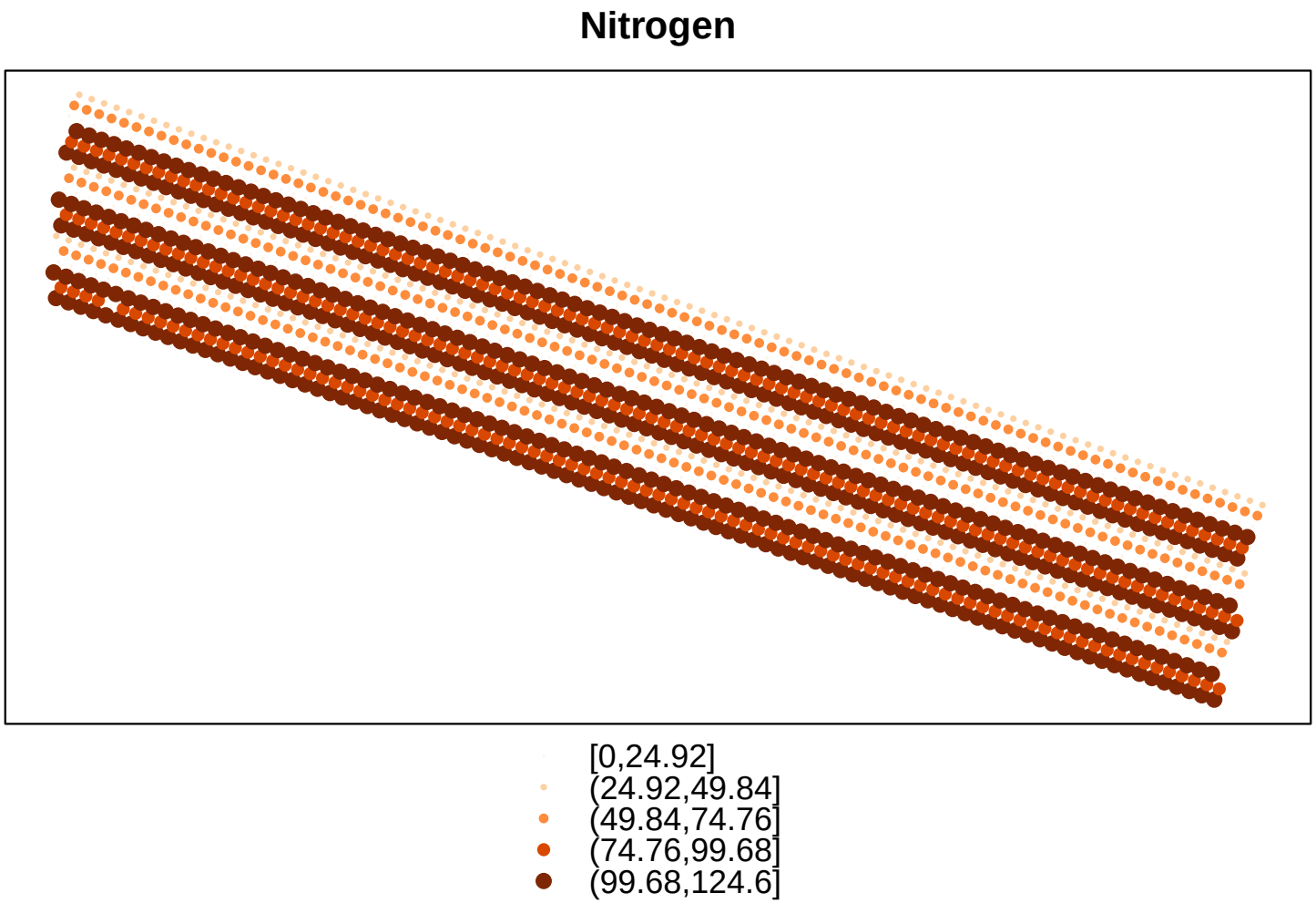
Objective

Do some of the explanatory variables influenced the observed yield variability in the field ?

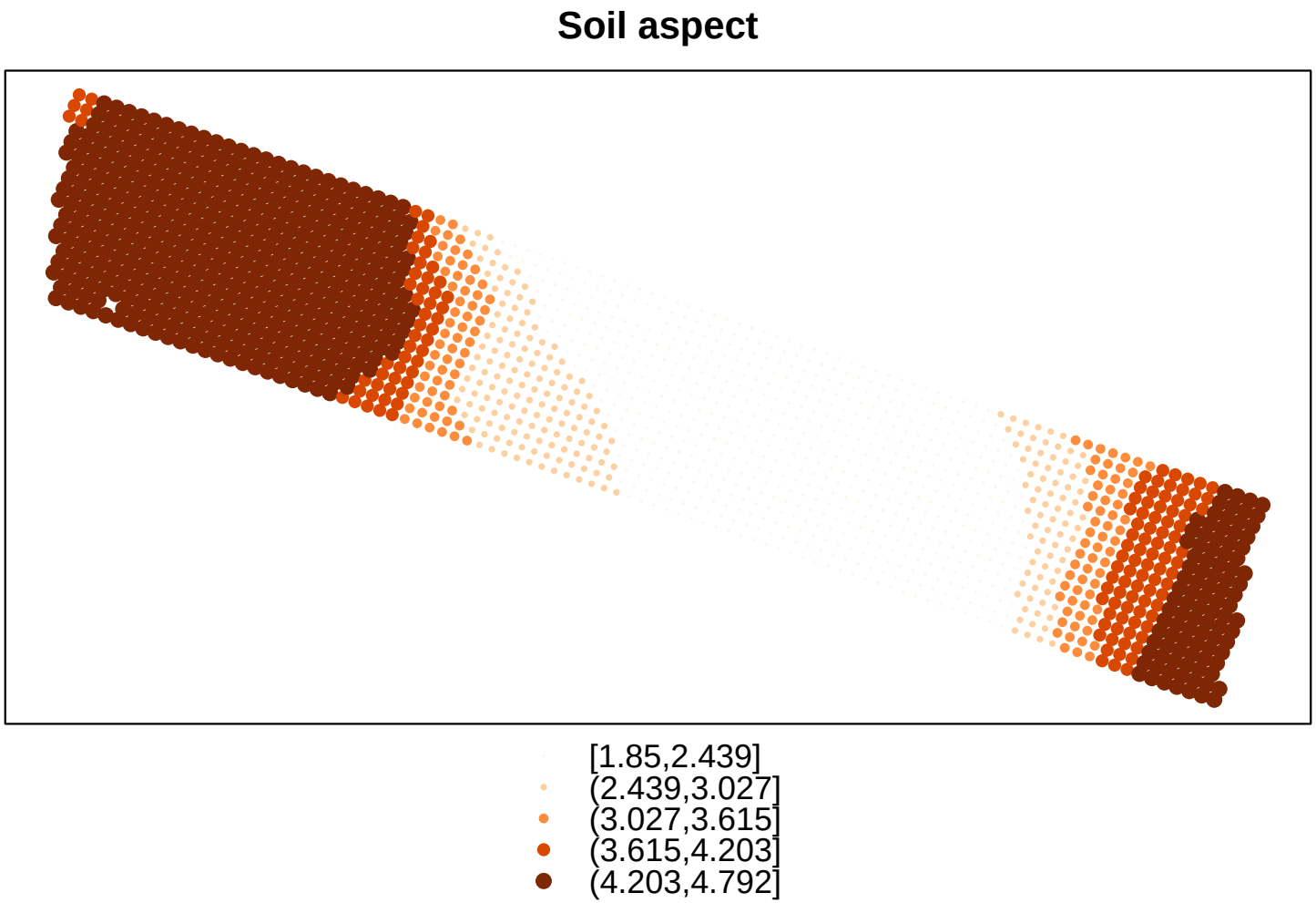
Example Las Rosas (2/16)



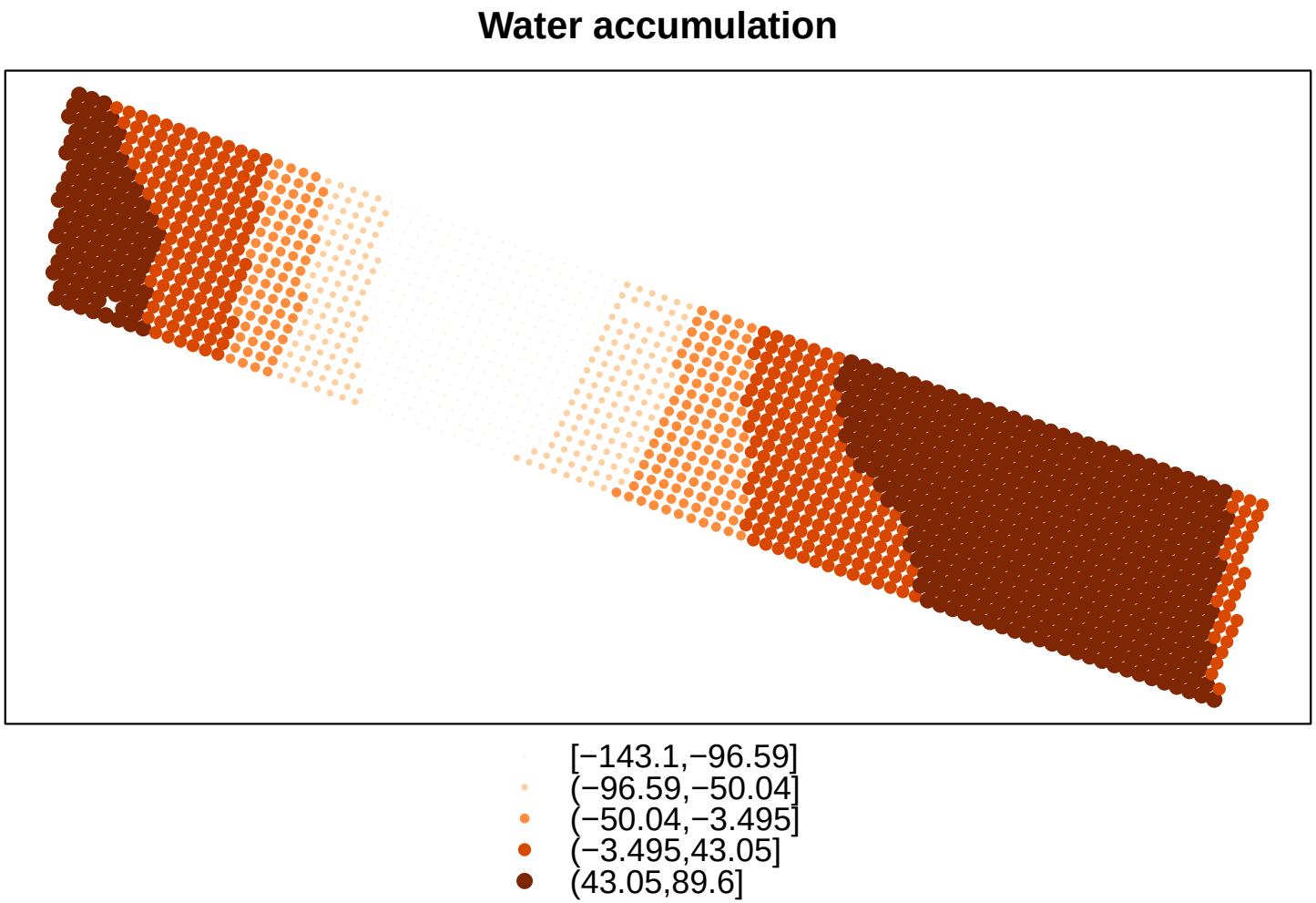
Example Las Rosas (3/16)



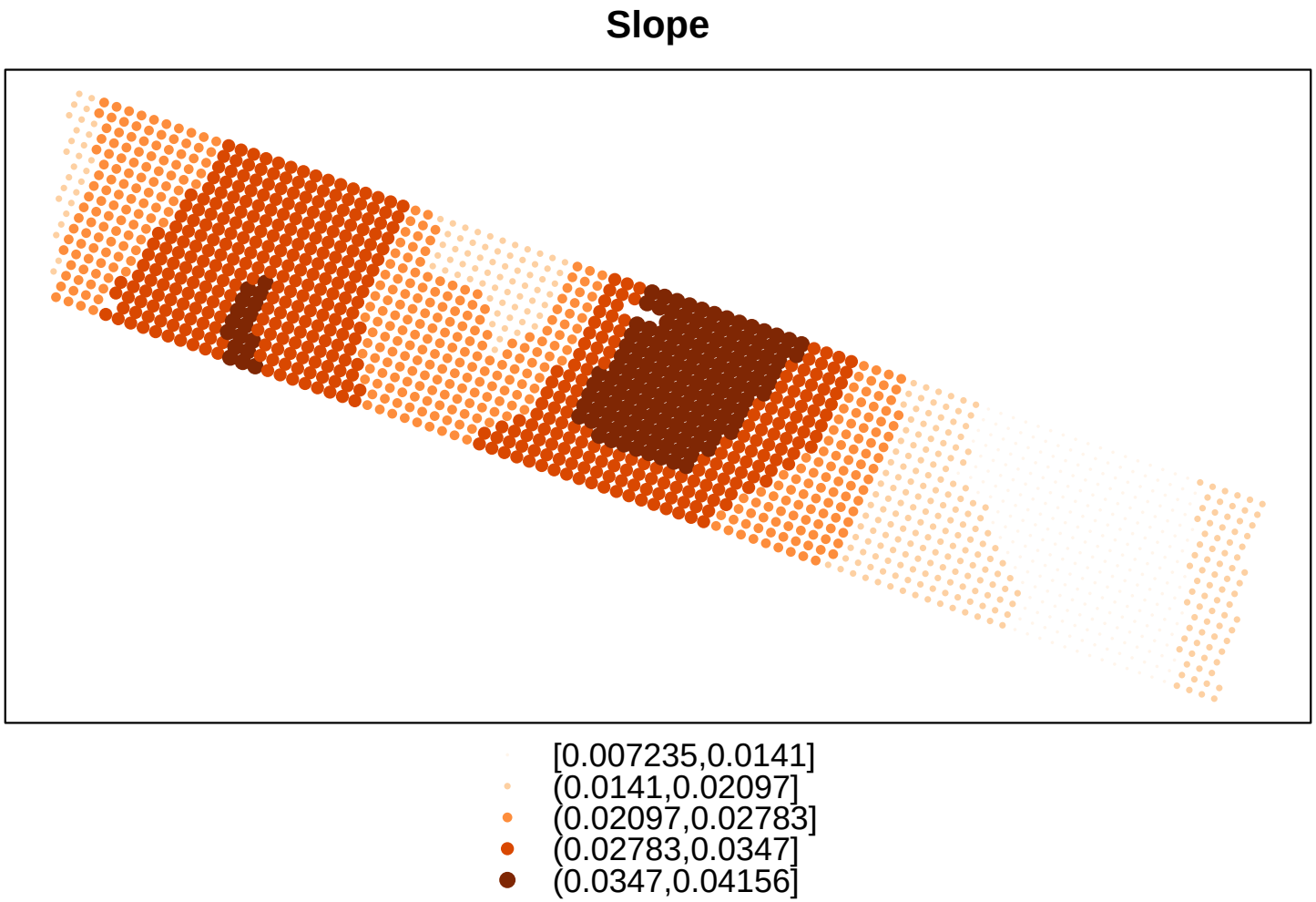
Example Las Rosas (4/16)



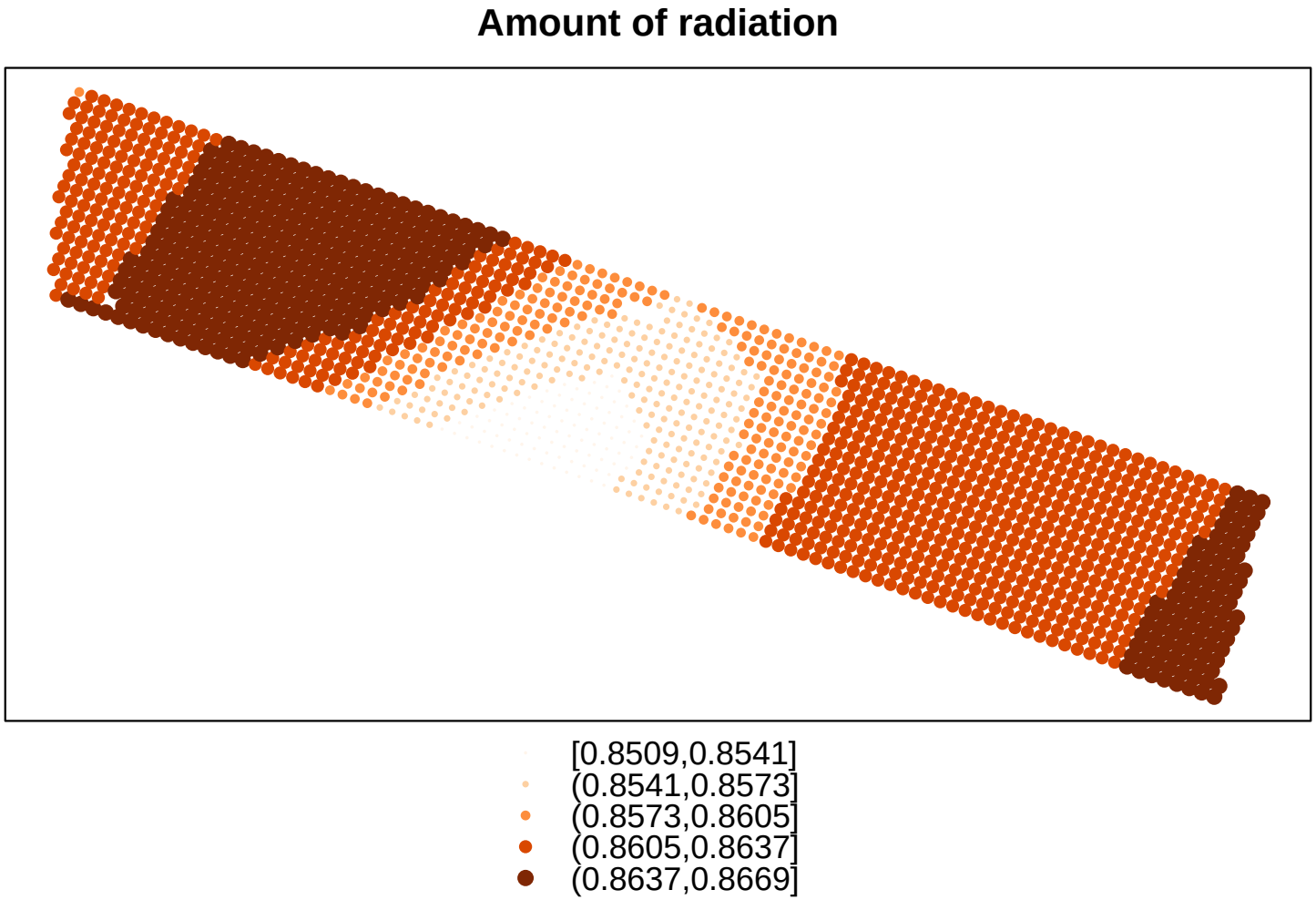
Example Las Rosas (5/16)



Example Las Rosas (6/16)



Example Las Rosas (7/16)



Example Las Rosas (8/16)

A linear regression model `model2.lm` has been proposed to explained the yield using all these explanatory variables

$$\begin{aligned} Yield_i &= \beta_0 + \beta_1 N_i + \beta_2 aspect_i + \beta_3 accu_i + \beta_4 accu_i \times slope_i + \beta_5 slope_i^2 \\ &\quad + \beta_6 accu_i \times hshade_i + \epsilon_i, \qquad (6) \\ \epsilon_i &\underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2). \end{aligned}$$

```
f<-as.formula("YIELD~N+aspect+accu+I(accu*slope)+I(slope^2)
               +I(accu*hshade)")
model2.lm<-lm(f,data=Xutm)
```

Assumptions should be checked.

Example Las Rosas (9/16)

```
drop1(model2.lm, . ~ ., test="F")

## Single term deletions
##
## Model:
## YIELD ~ N + aspect + accu + I(accu * slope) + I(slope^2)
##                  Df Sum of Sq       RSS   AIC F value    Pr(>F)
## <none>                        268414987 20404
## N                 1  15974896 284389884 20501  101.06 < 2.2e-16 ***
## aspect            1  60476672 328891660 20748  382.58 < 2.2e-16 ***
## accu              1  15952225 284367213 20501  100.91 < 2.2e-16 ***
## I(accu * slope)   1  38921239 307336226 20633  246.22 < 2.2e-16 ***
## I(slope^2)        1 107828744 376243732 20978  682.13 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example Las Rosas (10/16)

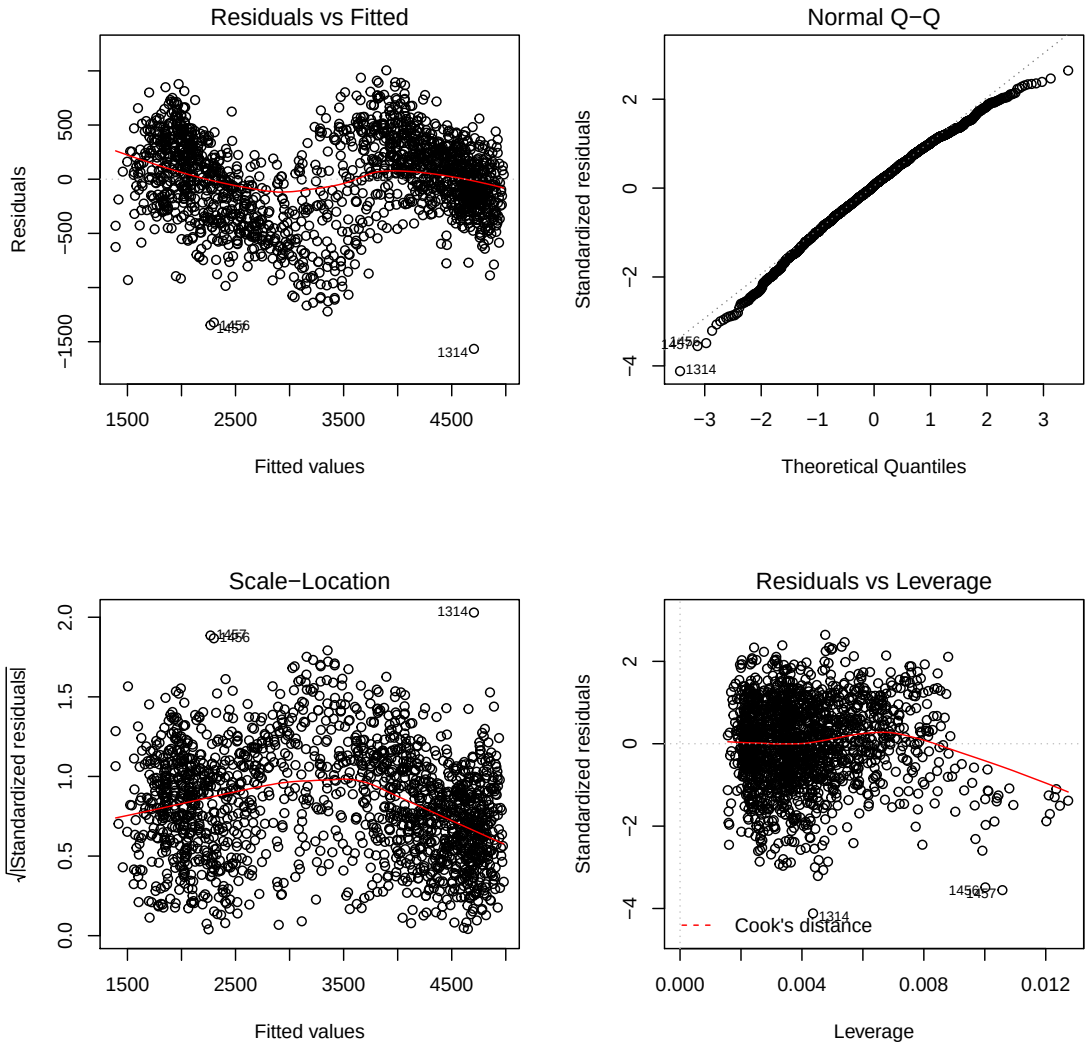


Figure 1 – Diagnostic plots for `model2.lm`.

Example Las Rosas (11/16)

```
ks.test(model2.lm$res, "pnorm", mean = 0, sd = sd(model2.lm$res))

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  model2.lm$res
## D = 0.032567, p-value = 0.05385
## alternative hypothesis: two-sided
```

Example Las Rosas (12/16)

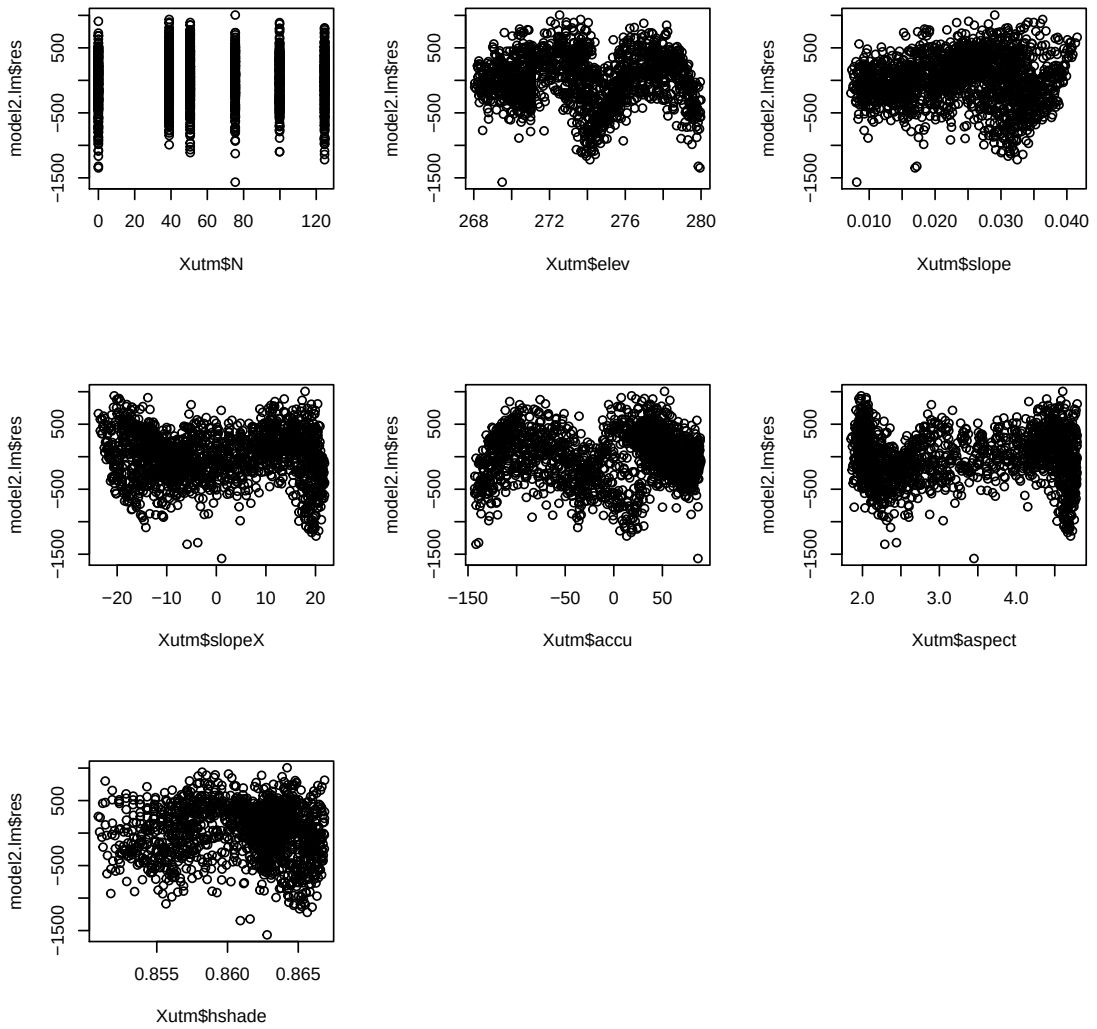


Figure 2 – Residuals of `model2.lm` against every possible explanatory variable.

Example Las Rosas (13/16)

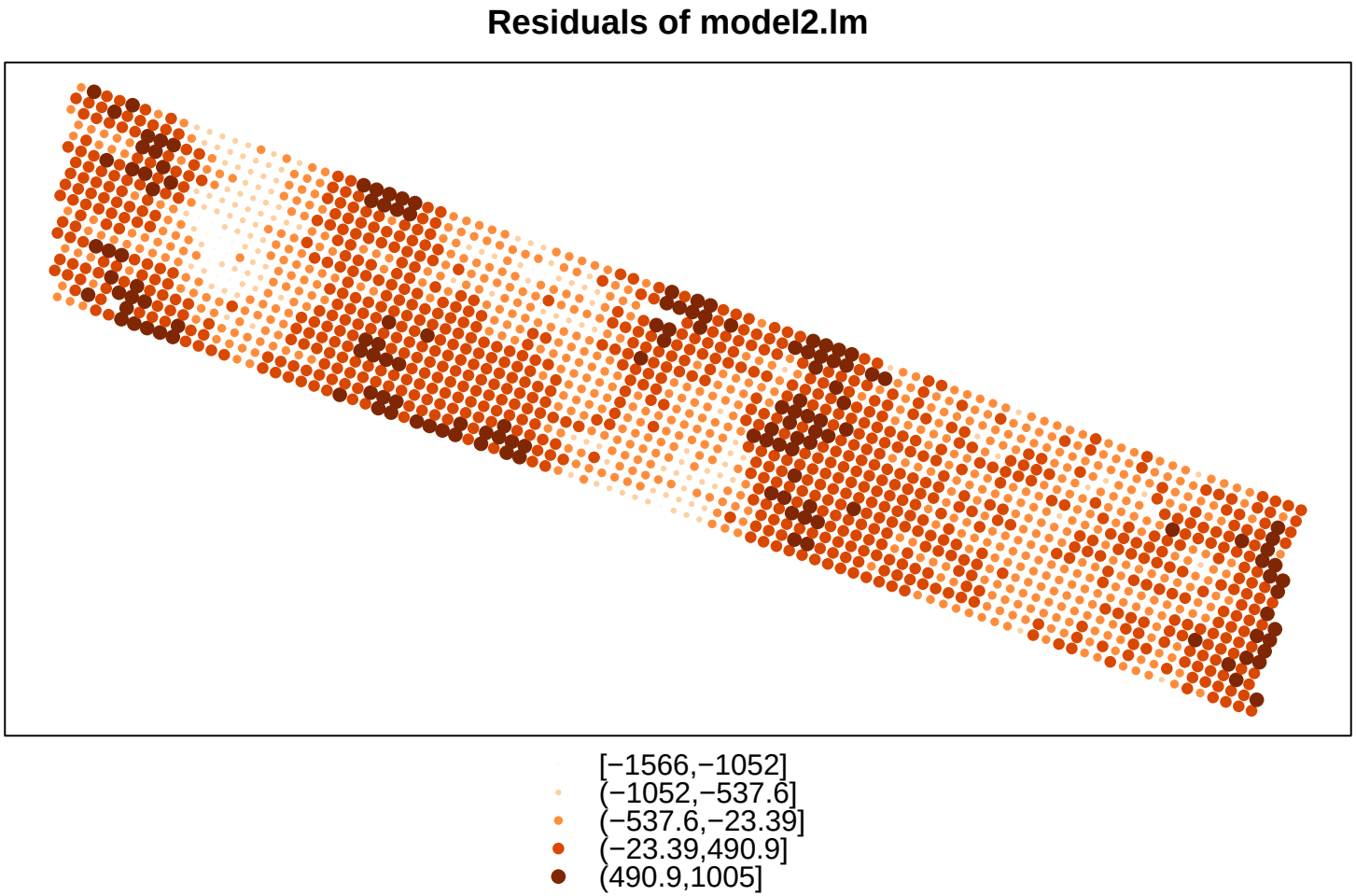


Figure 3 – Bubble map for residuals of `mod9`.

Example Las Rosas (14/16)

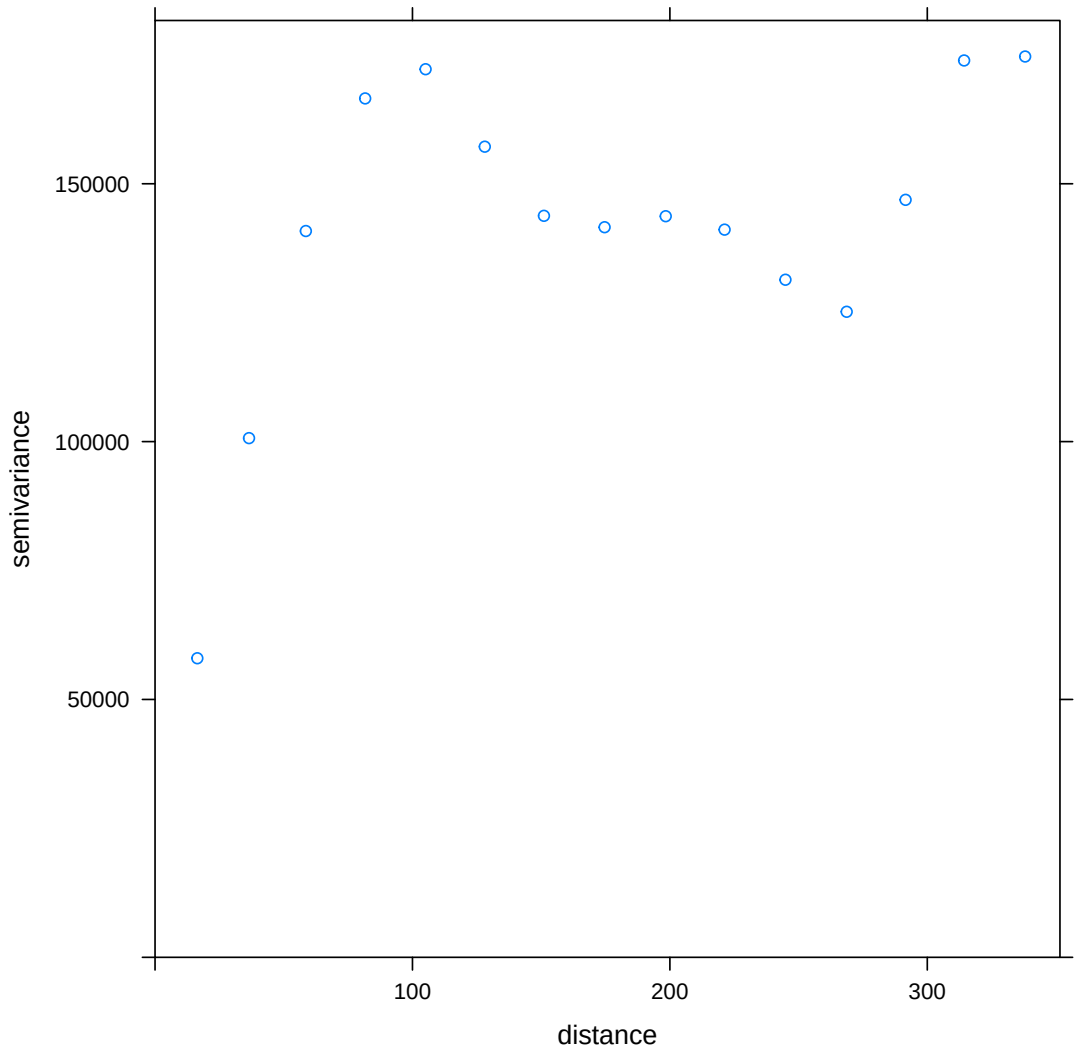


Figure 4 – Semi-variogram for the residuals of `model2.lm`.

Example Las Rosas (15/16)

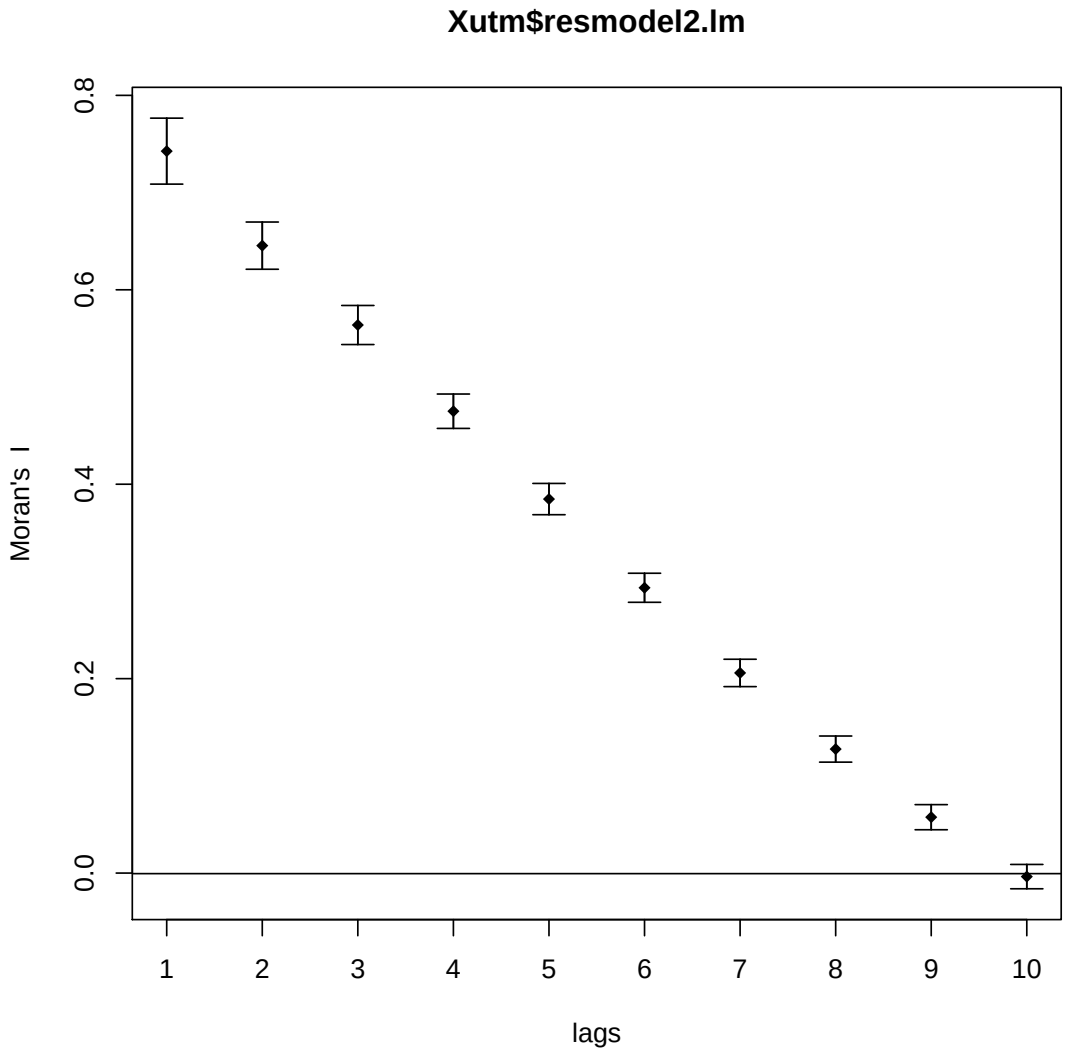


Figure 5 – Moran correlogram for residuals of `model2.lm`.

Example Las Rosas (16/16)

```
library(spdep)
nlist <- knn2nb(knearneigh(Xutm,k=8))
W <- nb2listw(nlist,style="W")
lm.morantest(model2.lm,W)

##
##  Global Moran I for regression residuals
##
## data:
## model: lm(formula = f, data = Xutm)
## weights: W
##
## Moran I statistic standard deviate = 60.822, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Observed Moran I      Expectation         Variance
##     0.7219463382    -0.0030524805     0.0001420855
```

Table of Contents

1 Sources and Consequences of Spatial Autocorrelation

2 Working example : Las Rosas

3 Spatial Lag Model
- Without explanatory variable
- With explanatory variable
- About the variance-covariance matrix of Y
- Fitting the model

4 Spatial Error Model

5 Choosing Between Spatial Lag and Spatial Error models

6 Extended Linear Models

7 Bibliography

Spatial lag model without explanatory variables

A spatial lag model with zero mean value and no explanatory variable has the form :

$$\begin{aligned} Y &= \rho WY + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma^2 I), \end{aligned} \tag{7}$$

where WY represents the spatial lag.

Interpretation

The value of Y at one location is directly associated with the values of the process Y at nearby locations.

For instance high productiovity of a plant at one location is associated with high productivity at nearby locations (but there is no notion of causality).

Spatial lag model with explanatory variables

A spatial lag model with explanatory variables :

$$\begin{aligned} Y &= \rho WY + X\beta + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma^2 I). \end{aligned} \tag{8}$$

Interpretation

This model can be interpreted using three points of view (Anselin 1992).

1. Specification of the spatial weights matrix W and estimation of ρ are indicators of the nature and strength of spatial interaction.
2. $Y = (I - \rho W)^{-1}(X\beta + \epsilon)$, and $\mathbb{E}(Y) = (I - \rho W)^{-1}X\beta$: non-linear effect of the spatial autocorrelation on the expected value of Y. The influence of the spatial structure is modelled through the error term and through the explanatory variables (influence of the neighborhood).

Prediction $\hat{Y} = (I - \hat{\rho} W)^{-1} X\hat{\beta}$ is mainly driven by the neighborhood. If we use $\hat{Y} = X\hat{\beta}$, we have a bias $-(\rho W)^{-1} X\beta$.

About the variance-covariance matrix of Y

$$var(Y) = \sigma^2(I - \rho W)^{-1}(I - \rho W')^{-1}.$$

- Impacted by the magnitude of the variance of the error term σ^2.
- Impacted by the spatial structure through the term $(I - \rho W)^{-1}(I - \rho W')^{-1}$.
- Enforced by the model, we do not have to specify it. **The spatial autocorrelation structure of Y is enforced by the model.**

Fitting the model

Estimation of the parameters β, σ^2 and ρ

Maximum likelihood approach.

In practice

The expressions of $\hat{\beta}$, $\hat{\sigma^2}$ and $\hat{\rho}$ that maximise the likelihood are not easy to obtain (it would be much easier if ρ was known)

$\Rightarrow$ use of a numerical scheme analogous to the Newton-Raphson method :

- A value of $\hat{\rho}$ is fixed.
- Maximum likelihood estimates $\hat{\beta}$ and $\hat{\sigma^2}$ calculated with $\hat{\rho}$ fixed.
- The two preceding steps are iterated : another value of $\hat{\rho}$ increasing the likelihood is fixed, $\hat{\beta}$ and $\hat{\sigma^2}$ are calculated to maximise the likelihood, then fix $\hat{\rho}$ again,...

Spatial lag model : Las Rosas (1/3)

```
library(spdep)
nlist <- knn2nb(knearneigh(Xutm,k=8))
W <- nb2listw(nlist,style="W")
Xutm$YIELD_scaled <- (Xutm$YIELD-mean(Xutm$YIELD))/sd(Xutm$YIELD)
Xutm$slope_scaled <- (Xutm$slope-mean(Xutm$slope))/sd(Xutm$slope)
f <- as.formula("YIELD_scaled ~ N + accu + aspect + I(accu*slope_scaled)
                +I(slope_scaled^2)+I(accu*hshade)")
mod.lag <- lagsarlm(f,data=Xutm,listw=W)
summary(mod.lag)
```

Spatial lag model : Las Rosas (2/3)

```
##
## Call:lagsarlm(formula = f, data = Xutm, listw = W)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.2674603 -0.1146204 -0.0013436  0.1172232  0.6719409
##
## Type: lag
## Coefficients: (asymptotic standard errors)
##                           Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)             -3.6686e-02  1.7260e-02 -2.1255 0.0335489
## N                        1.5924e-03  1.1088e-04 14.3611 < 2.2e-16
## accu                     1.2602e-03  1.6629e-04  7.5785 3.486e-14
## aspect                  -1.5240e-02  4.4662e-03 -3.4122 0.0006444
## I(accu * slope_scaled)   2.9292e-04  9.2681e-05  3.1606 0.0015747
```

Spatial lag model : Las Rosas (3/3)

```
## Rho: 0.90687, LR test value: 2405.7, p-value: < 2.22e-16
## Asymptotic standard error: 0.011712
##     z-value: 77.43, p-value: < 2.22e-16
## Wald statistic: 5995.4, p-value: < 2.22e-16
##
## Log likelihood: 327.3626 for lag model
## ML residual variance (sigma squared): 0.033407, (sigma: 0.18278)
## Number of observations: 1704
## Number of parameters estimated: 7
## AIC: -640.73, (AIC for lm: 1762.9)
## LM test for residual autocorrelation
## test value: 42.144, p-value: 8.4802e-11
```

Table of Contents

1 Sources and Consequences of Spatial Autocorrelation

2 Working example : Las Rosas

3 Spatial Lag Model

4 Spatial Error Model
- Formulation
- About the variance-covariance matrix of Y and fitting

5 Choosing Between Spatial Lag and Spatial Error models

6 Extended Linear Models

7 Bibliography

Spatial Error Model (1/2)

$$\begin{aligned} Y &= X\beta + \eta \\ \eta &= \lambda W\eta + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma^2 I). \end{aligned} \tag{9}$$

Interpretation

1. Like a classical linear model, but with a correlated structure for the error term.
   This autocorrelation is generally considered to be a nuisance : the primary interest is often the relationship between the explanatory variables X and the response variable Y.
   The spatial autocorrelation is just taken into account through the error term.

Spatial Error Model (2/2)

2. The influence of the spatial structure is modelled only on the error term $Y = X\beta + (I - \lambda W)^{-1}\epsilon$.

Prediction $\hat{Y} = X\hat{\beta}$ is driven by the values of the explanatory variables at the location for which we want the prediction. Be careful, to have an unbiased estimation of β, you must use the spatial error model and not the classical linear model if your data are driven by this spatial error model.

This model can be written as a classical linear model :

$$\begin{aligned} Y - \lambda WY &= X\beta - \lambda WY + \eta \\ &= (X - \lambda WX)\beta + \epsilon \\ \tilde{Y} &= \tilde{X}\beta + \epsilon \end{aligned}$$

About the variance-covariance matrix of Y and fitting

$$var(Y) = \sigma^2 (I - \lambda W)^{-1} (I - \lambda W')^{-1}. \quad (10)$$

- Impacted by the magnitude of the variance of the error term σ^2.
- Impacted by the spatial structure through the term $(I - \lambda W)^{-1}(I - \lambda W')^{-1}$.
- Enforced by the model, we do not have to specify it. **The spatial autocorrelation structure of Y is enforced by the model.**

Fitting the model

The approach is the same as for the spatial lag model, with ρ replaced by λ.

Spatial error model : Las Rosas (1/2)

```
mod.err <- errorsarlm(f,data=Xutm,listw=W)
summary(mod.err)
```

```
##
## Call:errorsarlm(formula = f, data = Xutm, listw = W)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.2112753 -0.1053701  0.0024866  0.1097140  0.6707100
##
## Type: error
## Coefficients: (asymptotic standard errors)
##                          Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)             0.48112487  0.16667563  2.8866  0.003894
## N                       0.00195918  0.00011087 17.6711 < 2.2e-16
## accu                    0.01252811  0.00080061 15.6483 < 2.2e-16
## aspect                 -0.14706321  0.04800230 -3.0637  0.002186
## I(accu * slope_scaled)  0.00198320  0.00077652  2.5540  0.010651
```

Spatial error model : Las Rosas (2/2)

```
## Lambda: 0.92271, LR test value: 2482.1, p-value: < 2.22e-16
## Asymptotic standard error: 0.011007
##     z-value: 83.826, p-value: < 2.22e-16
## Wald statistic: 7026.9, p-value: < 2.22e-16
##
## Log likelihood: 365.5709 for error model
## ML residual variance (sigma squared): 0.031578, (sigma: 0.1777)
## Number of observations: 1704
## Number of parameters estimated: 7
## AIC: -717.14, (AIC for lm: 1762.9)
```

Table of Contents

1 Sources and Consequences of Spatial Autocorrelation

2 Working example : Las Rosas

3 Spatial Lag Model

4 Spatial Error Model

5 Choosing Between Spatial Lag and Spatial Error models

6 Extended Linear Models

7 Bibliography

Choosing Between Spatial Lag and Spatial Error models (1/2)

Spatial autocorrelation detected in residuals of a classical linear model
$\Rightarrow$ take into account this autocorrelation
$\Rightarrow$ choose between spatial lag model and spatial error model (or an extended linear model).

These two models can be combined in a single model of the form :

$$\begin{aligned} Y &= \rho W_1 Y + X\beta + \eta \\ \eta &= \lambda W_2 \eta + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma^2 I), \end{aligned} \tag{11}$$

where W_1 cannot be equal to W_2 or X cannot simply be a vector of ones (for identifiability).

Choosing Between Spatial Lag and Spatial Error models (2/2)

The problem with choosing between the two models can be expressed as two hypothesis tests :

$$\text{1st test} : \begin{cases} H_0 : & \rho = 0, \\ H_1 : & \rho \neq 0 \end{cases} \quad \text{and} \quad \text{2nd test} : \begin{cases} H_0 : & \lambda = 0, \\ H_1 : & \lambda \neq 0 \end{cases}$$

- H_0 kept for both tests $\Rightarrow$ keep a classical linear model, there is no spatial autocorrelation of the residuals.
- H_0 kept for the first test and H_1 non-rejected for the second test $\Rightarrow$ spatial error model.
- H_0 kept for the second test and H_1 non-rejected for the first test $\Rightarrow$ spatial lag model.

In practice :

- These tests are carried using the Lagrange multiplier test.
- Another possibility : use the AIC criteria.

Example Las Rosas (1/16)

```
model2.lm_scaled <- lm(YIELD_scaled ~ N + accu + aspect + I(accu*slope_scal
                       +I(slope_scaled^2)+I(accu*hshade), data=Xutm)
lm.LMtests(model2.lm_scaled,W,test=c("LMlag","LMerr","SARMA"))
```

```
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = YIELD_scaled ~ N + accu + aspect + I(accu *
## slope_scaled) + I(slope_scaled^2) + I(accu * hshade), data = Xutm)
## weights: W
##
## LMlag = 3696.2, df = 1, p-value < 2.2e-16
```

Example Las Rosas (2/16)

```
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = YIELD_scaled ~ N + accu + aspect + I(accu *
## slope_scaled) + I(slope_scaled^2) + I(accu * hshade), data = Xutm)
## weights: W
##
## LMerr = 3893.5, df = 1, p-value < 2.2e-16
##
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = YIELD_scaled ~ N + accu + aspect + I(accu *
## slope_scaled) + I(slope_scaled^2) + I(accu * hshade), data = Xutm)
## weights: W
##
## SARMA = 3898.9, df = 2, p-value < 2.2e-16
```

Example Las Rosas (3/16)

AIC $\Rightarrow$ we prefer the spatial error model.

The yield at one location is mainly driven by the values of the explanatory variables at this location.

If we assume interaction (competition) between corn plants, we should prefer the spatial lag model.

Example Las Rosas (4/16)

Improve the spatial error model by performing model selection.

- Try to remove explanatory variables or interactions between them and to include variables which are not present in `mod.err`.
- AIC criteria to select the best model.
- We remove the interaction `accu*slope_scaled`, and we include the variable `elev` which has been scaled.

```
Xutm$elev_scaled <- (Xutm$elev-mean(Xutm$elev))/sd(Xutm$elev)
f <- as.formula("YIELD_scaled ~ N + accu + aspect +I(slope_scaled^2)+
                 I(accu*hshade)+elev_scaled")
mod.err8 <- errorsarlm(f,data=Xutm,listw=W)
summary(mod.err8)
```

Example Las Rosas (5/16)

```
##
## Call:errorsarlm(formula = f, data = Xutm, listw = W)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.2008215 -0.1072744  0.0035153  0.1118625  0.6940611
##
## Type: error
## Coefficients: (asymptotic standard errors)
##                     Estimate Std. Error z value  Pr(>|z|)
## (Intercept)        0.6597909  0.1517124  4.3490 1.368e-05
## N                  0.0019562  0.0001116 17.5294 < 2.2e-16
## accu               0.5553112  0.1388971  3.9980 6.388e-05
## aspect            -0.1716027  0.0409821 -4.1873 2.823e-05
## I(slope_scaled^2) -0.1272388  0.0357075 -3.5634 0.0003661
## I(accu * hshade)  -0.6360292  0.1619698 -3.9268 8.607e-05
## elev_scaled       -0.3653518  0.1443791 -2.5305 0.0113899
```

Example Las Rosas (6/16)

```
##
## Lambda: 0.90054, LR test value: 1976, p-value: < 2.22e-16
## Asymptotic standard error: 0.012869
##     z-value: 69.977, p-value: < 2.22e-16
## Wald statistic: 4896.8, p-value: < 2.22e-16
##
## Log likelihood: 373.6497 for error model
## ML residual variance (sigma squared): 0.031777, (sigma: 0.17826)
## Number of observations: 1704
## Number of parameters estimated: 9
## AIC: -729.3, (AIC for lm: 1244.7)
```

Example Las Rosas (7/16)

We check that the assumptions are verified on the residuals of `mod.err8`.

```
ks.test(mod.err8$residuals,"pnorm", mean=0, sd=sd(mod.err8$residuals))

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  mod.err8$residuals
## D = 0.029614, p-value = 0.1007
## alternative hypothesis: two-sided
```

Example Las Rosas (8/16)

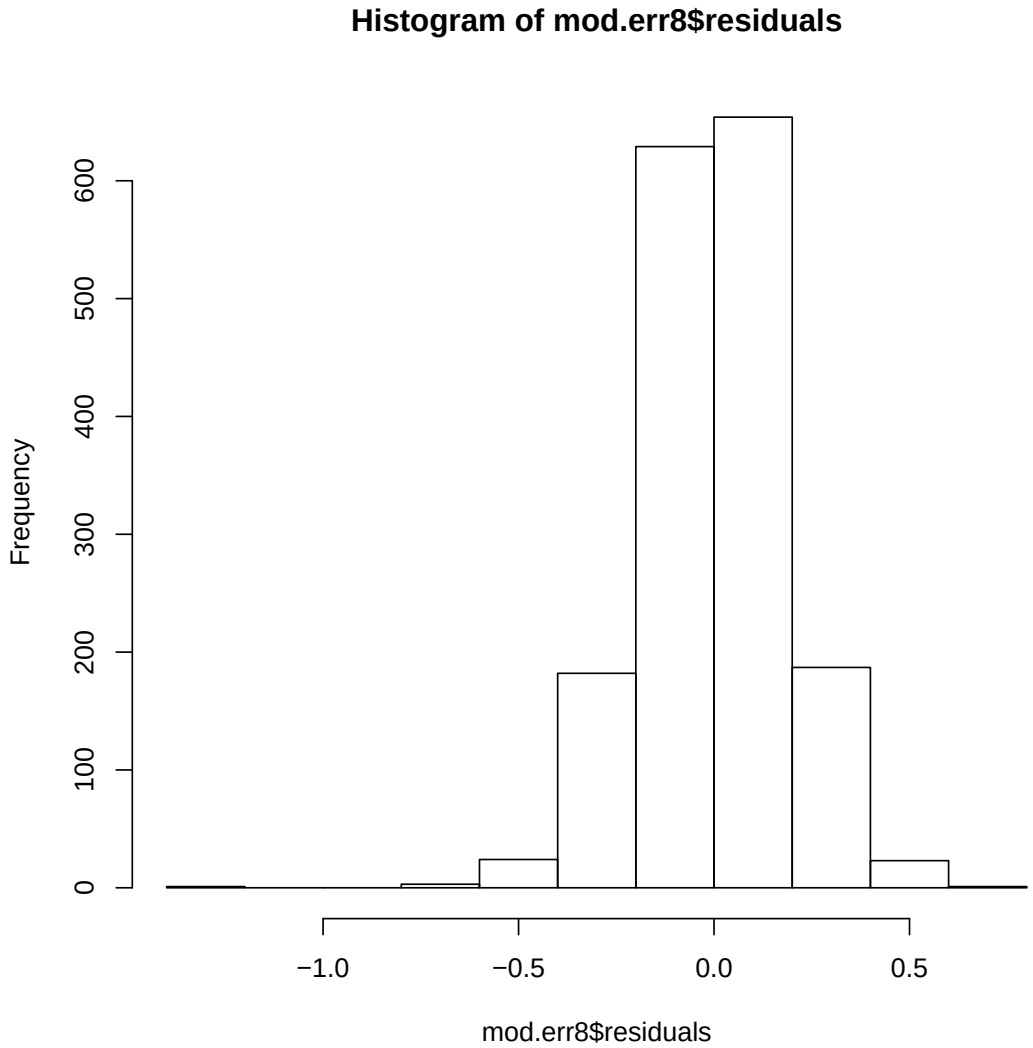


Figure 6 – Histogram of residuals of `mod.err8`.

Example Las Rosas (9/16)

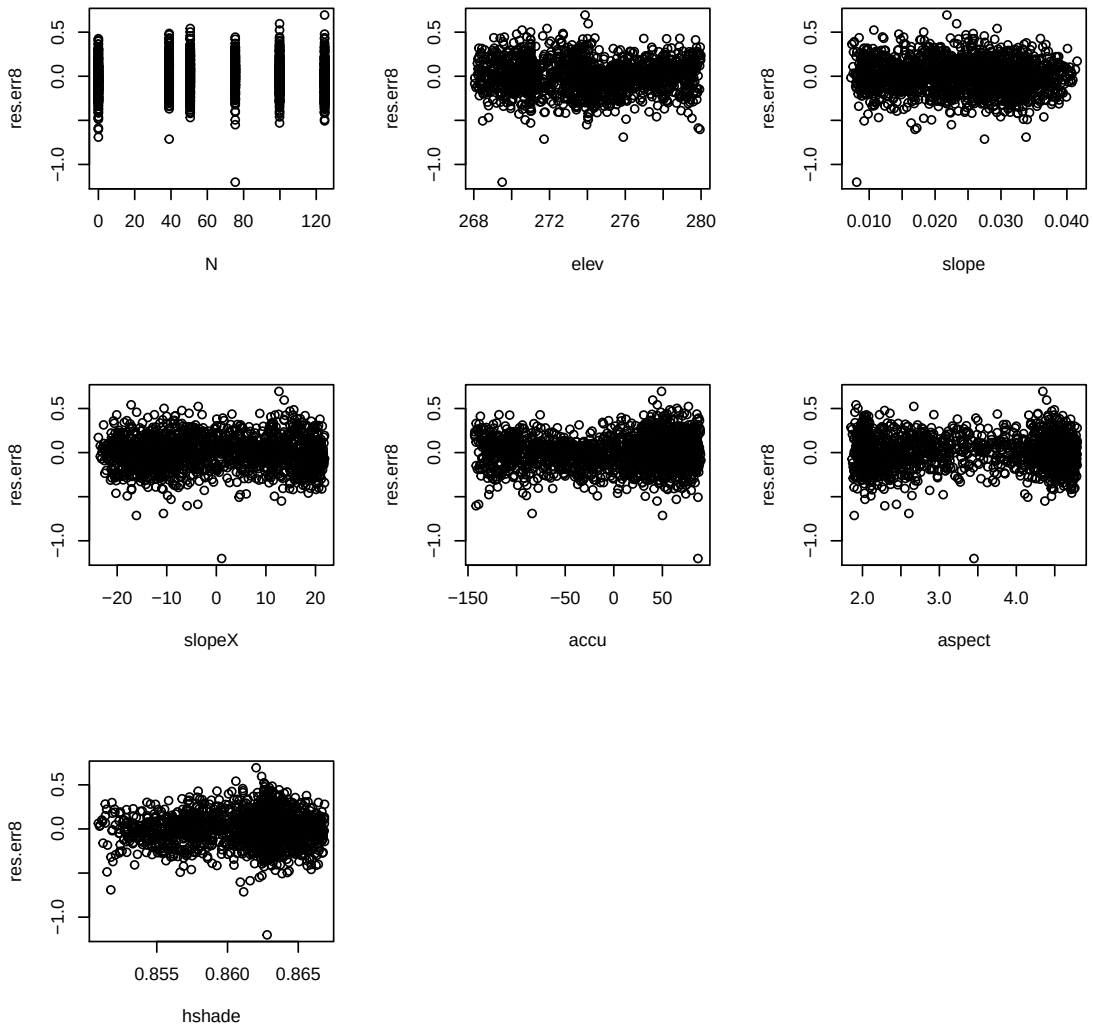


Figure 7 – Residuals of `mod.err8` against every possible explanatory variable.

Example Las Rosas (10/16)

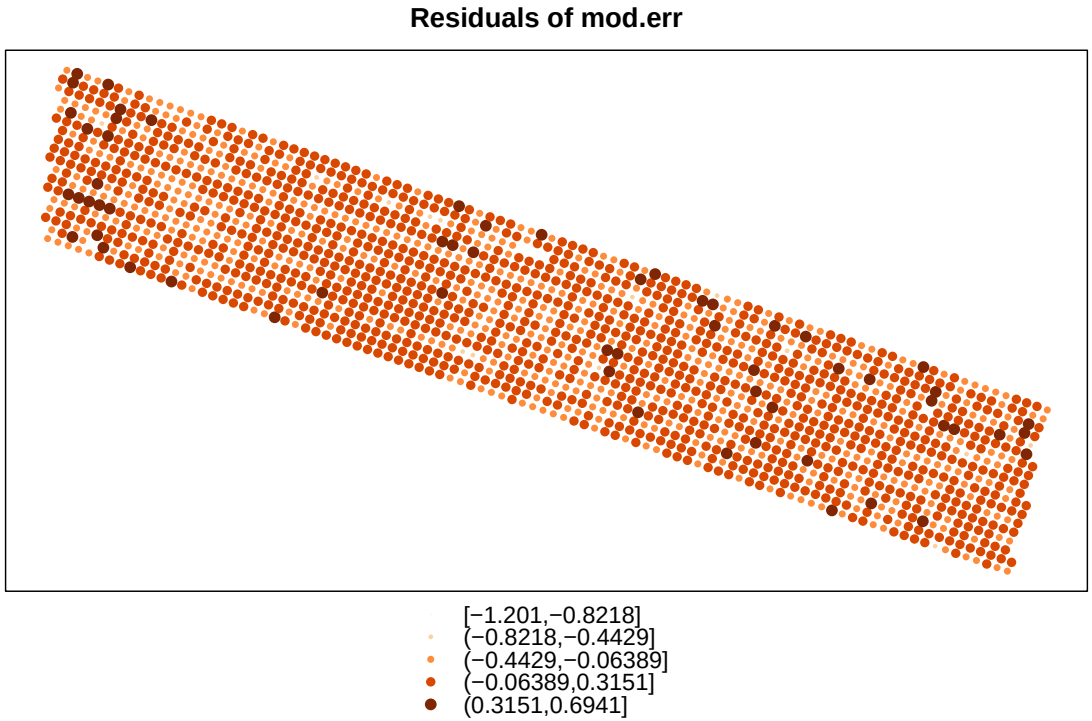


Figure 8 – Bubble map for residuals of `mod.err8`.

Example Las Rosas (11/16)

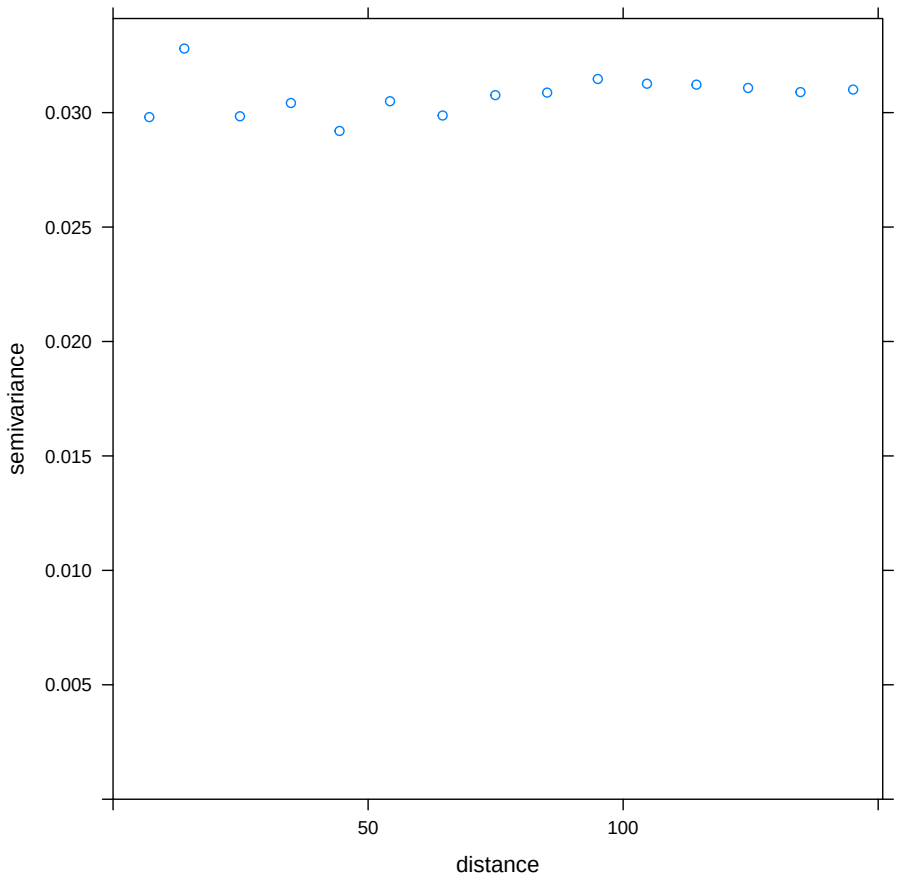


Figure 9 – Semi-variogram for the residuals of `mod.err8`.

Example Las Rosas (12/16)

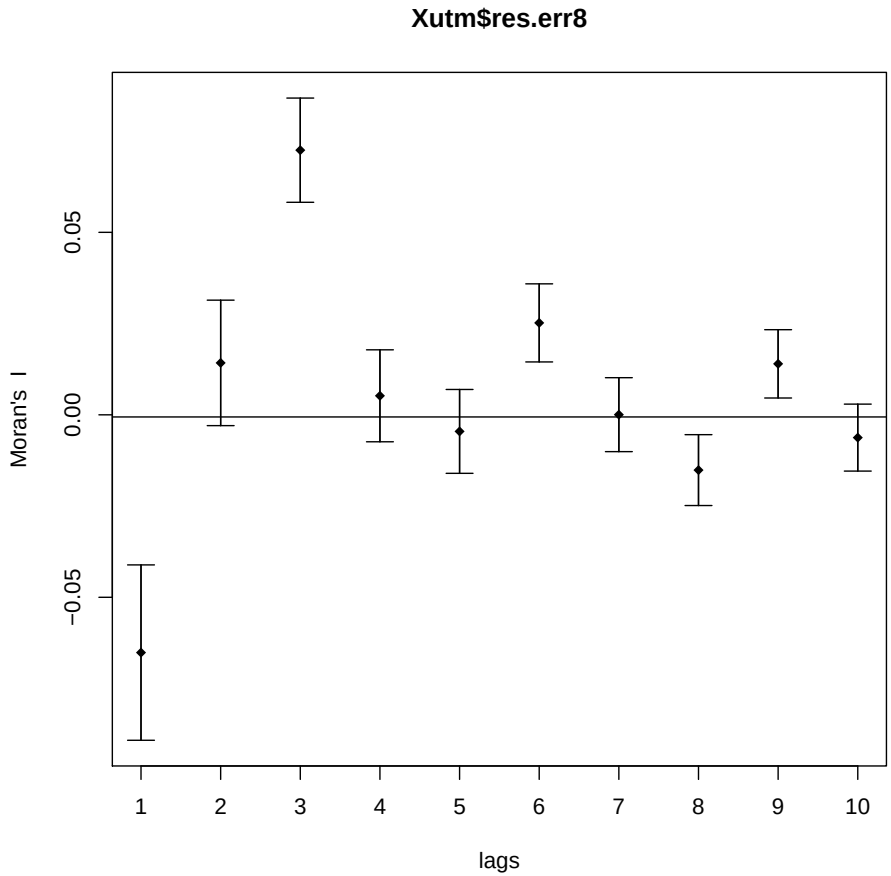


Figure 10 – Moran correlogram for residuals of `mod.err8`.

Example Las Rosas (13/16)

```
moran.mc(Xutm@data$res.err8,W,nsim=1000,alternative="greater")

##
##  Monte-Carlo simulation of Moran I
##
## data:  Xutm@data$res.err8
## weights: W
## number of simulations + 1: 1001
##
## statistic = -0.065154, observed rank = 1, p-value = 0.999
## alternative hypothesis: greater
```

Example Las Rosas (14/16)

```
pred <- as.data.frame(predict.sarlm(mod.err8))
head(pred)
```

```
##         fit     trend    signal
## 1 0.9903291 0.8177455 0.1725836
## 2 0.9004259 0.7922841 0.1081418
## 3 0.8699619 0.7556189 0.1143431
## 4 0.9720662 0.7043677 0.2676985
## 5 1.0043946 0.6304516 0.3739430
## 6 0.9097421 0.5436834 0.3660587
```

Example Las Rosas (15/16)

```
Xutm2 <- Xutm
Xutm2@data$N <- Xutm@data$N + 1
newpred <- as.data.frame(predict.sarlm(mod.err8, newdata=Xutm2, listw = W))
head(newpred)
```

```
##         fit     trend signal
## 1 0.8197017 0.8197017      0
## 2 0.7942404 0.7942404      0
## 3 0.7575751 0.7575751      0
## 4 0.7063240 0.7063240      0
## 5 0.6324078 0.6324078      0
## 6 0.5456396 0.5456396      0
```

```
diff <- newpred$fit-pred$fit
Xutm@data$diff <- diff
spplot(Xutm, "diff", col.regions=brewer.pal(9,"Oranges"),
            cex=.2*(1:5), aspect=1/2, main="Predicted differences of yield
```

Example Las Rosas (16/16)

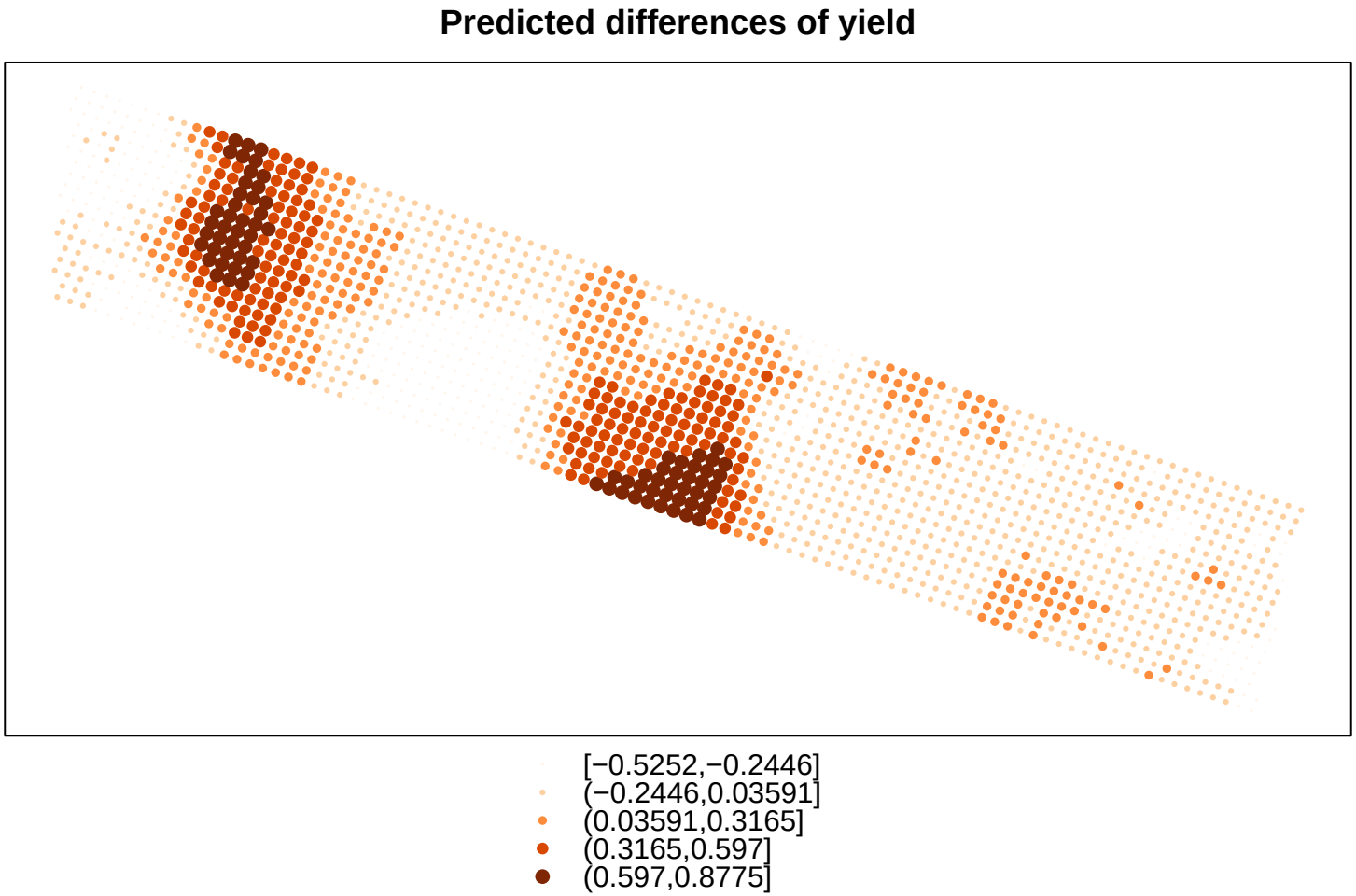


Figure 11 – Bubble map for the predicted differences of yield when N is increased

Table of Contents

1 Sources and Consequences of Spatial Autocorrelation

2 Working example : Las Rosas

3 Spatial Lag Model

4 Spatial Error Model

5 Choosing Between Spatial Lag and Spatial Error models

6 Extended Linear Models
- Classical Linear Model versus Extended Linear Model
- Modelling Spatial Correlation

7 Bibliography

Classical linear model versus extended linear model (1/2)

Y quantitative variable to explain, explanatory variables quantitative or qualitative :

$$Y = X\beta + \epsilon \qquad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I). \tag{12}$$

Possible extensions

On the variance-covariance matrix of the residuals (among others).

In classical linear models :

- The residuals (therefore the observations) are supposed independent.
- The residuals are supposed homoscedastic.

Classical linear model versus extended linear model (2/2)

$$Y = X\beta + \epsilon \qquad \text{with} \quad \epsilon \sim \mathcal{N}(0, \Lambda). \tag{13}$$

1. If Λ is diagonal, but with varying coefficients on the diagonal, $\Rightarrow$ heteroscedasticity.
2. If Λ has non-null coefficients outside the diagonal $\Rightarrow$ correlation between the residuals, dependence structure of the residuals. This dependence can be temporal, spatial or more general.

In practice, once a modelisation has been chosen

- Parameters of these extended linear models (regression coefficients and coefficients of the variance-covariance matrix) estimated using the maximum likelihood estimators.
- Numerically obtained by solving an ordinary least-squares problem.

Modelling Spatial Correlation

Extended linear model vs regression models for spatially autocorrelated data

- Models for spatially autocorrelated data : special cases of extended linear models.
- In extended linear models Λ can take any form.
- In the regression models designed for spatially autocorrelated data, the form of Λ is enforced by the model.
- Regression models for spatially autocorrelated data often more intuitive than extended linear models.

Choosing the modelisation of the spatial dependency (1/2)

Look at the form of the semi-variogram

- Choosing the form of the variance-covariance matrix $\Lambda \Leftrightarrow$ to choose a semi-variogram pattern.
- The form of the empirical semi-variogram can guide us to choose a semi-variogram pattern.

Model	Formula (ρ the range)
Exponential	$\gamma(d,\rho) = 1 - \exp(-d/\rho)$
Gaussian	$\gamma(d,\rho) = 1 - \exp\left[-(d/\rho)^2\right]$
Linear	$\gamma(d,\rho) = 1 - (1 - d/\rho)\mathbb{1}(d < \rho)$
Rational quadratic	$\gamma(d,\rho) = (d/\rho)^2 / \left[1 + (d/\rho)^2\right]$
Spherical	$\gamma(d,\rho) = 1 - \left[1 - 1.5(d/\rho) + 0.5(d/\rho)^3\right]\mathbb{1}(d < \rho)$

Table 1 – Some isotropic semi-variogram models for spatial correlation structures.

Use classical model selection methods

AIC, BIC, tests between nested models.

Choosing the modelisation of the spatial dependency (2/2)

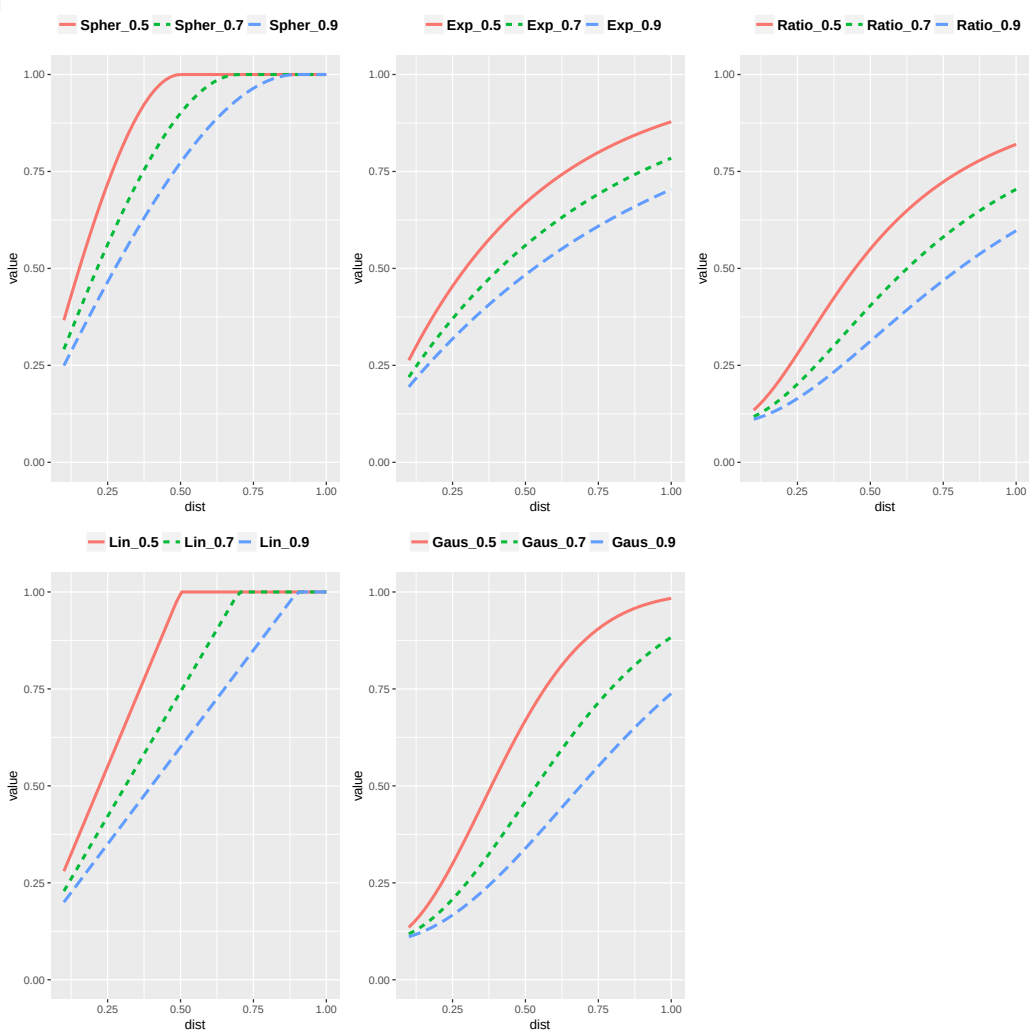


Figure 12 – Different semi-variogram patterns : Spherical, Exponential, Rational quadratic, Linear and Gaussian. Each pattern has a nugget of 0.1. The value of

Example Las Rosas (1/7)

```
library(nlme)
model2.lm <- gls(YIELD_scaled ~ N + accu + aspect + I(accu*slope_scaled)
                 +I(slope_scaled^2)+I(accu*hshade), data=Xutm)
modSpher <- gls(YIELD_scaled ~ N + accu + aspect + I(accu*slope_scaled)
                +I(slope_scaled^2)+I(accu*hshade), data=Xutm,
                correlation=corSpher(form=~x+y,nugget=T))
modLin <- gls(YIELD_scaled ~ N + accu + aspect + I(accu*slope_scaled)
              +I(slope_scaled^2)+I(accu*hshade), data=Xutm,
              correlation=corLin(form=~x+y,nugget=T))
modRatio <- gls(YIELD_scaled ~ N + accu + aspect + I(accu*slope_scaled)
                +I(slope_scaled^2)+I(accu*hshade), data=Xutm,
                correlation=corRatio(form=~x+y,nugget=T))
modGaus <- gls(YIELD_scaled ~ N + accu + aspect + I(accu*slope_scaled)
               +I(slope_scaled^2)+I(accu*hshade), data=Xutm,
               correlation=corGaus(form=~x+y,nugget=T))
modExp <- gls(YIELD_scaled ~ N + accu + aspect + I(accu*slope_scaled)
              +I(slope_scaled^2)+I(accu*hshade), data=Xutm,
              correlation=corExp(form=~x+y,nugget=T))
```

Example Las Rosas (2/7)

```
AIC(modSpher,modLin,modRatio,modGaus,modExp)

##           df       AIC
## modSpher 10 -837.3943
## modLin   10 -830.0858
## modRatio 10 -783.4707
## modGaus  10 -761.8102
## modExp   10 -832.0582

anova(model2.lm,modSpher)

##            Model df       AIC       BIC    logLik   Test  L.Ratio p-value
## model2.lm      1  8 1463.0420 1506.5349 -723.5210
## modSpher       2 10 -837.3943 -783.0281  428.6972 1 vs 2 2304.436  <.0001
```

```
VarioSpher_raw <- Variogram(modSpher, form =~ LONGITUDE + LATITUDE,
                            robust = TRUE, maxDist = 350, resType = "pearson")
VarioSpher_normalized <- Variogram(modSpher, form =~ LONGITUDE + LATITUDE,
                              robust = TRUE, maxDist = 350, resType = "normalized
```

Example Las Rosas (3/7)

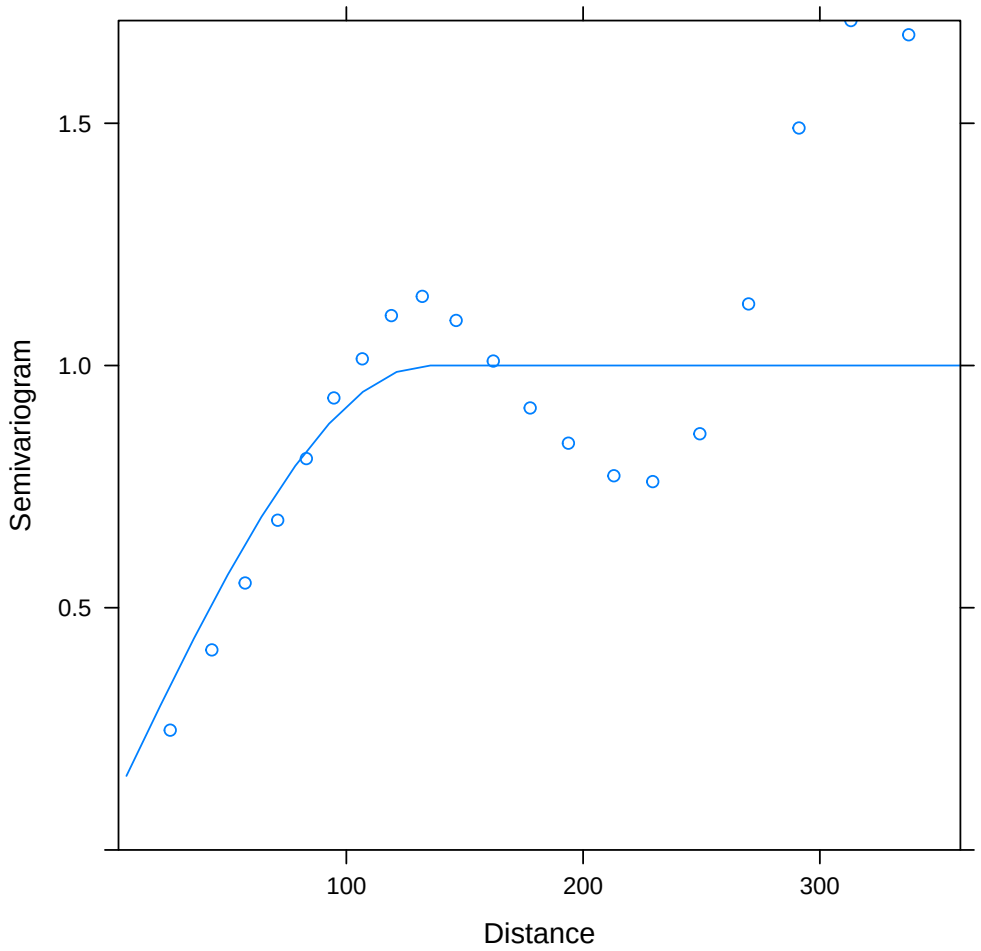


Figure 13 – Semi-variogram for the raw residuals of `modSpher`.

Example Las Rosas (4/7)

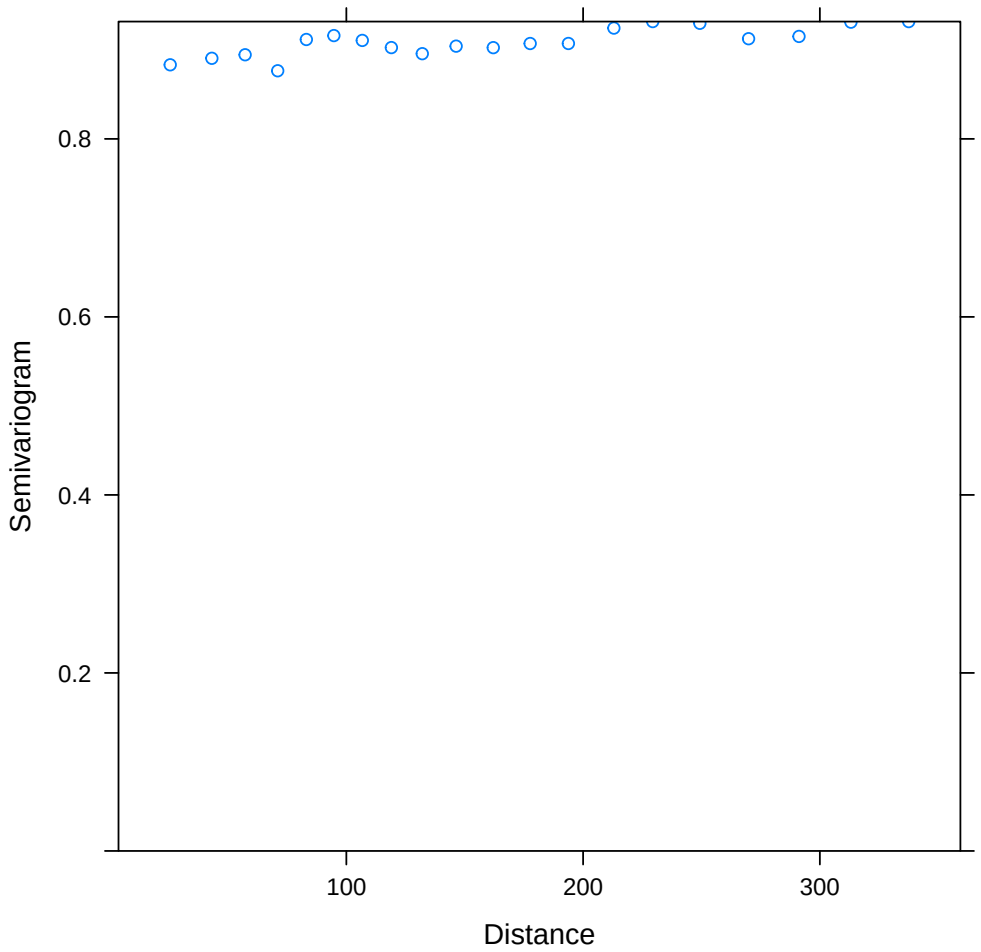


Figure 14 – Semi-variogram for the studentized residuals of `modSpher`.

Example Las Rosas(5/7)

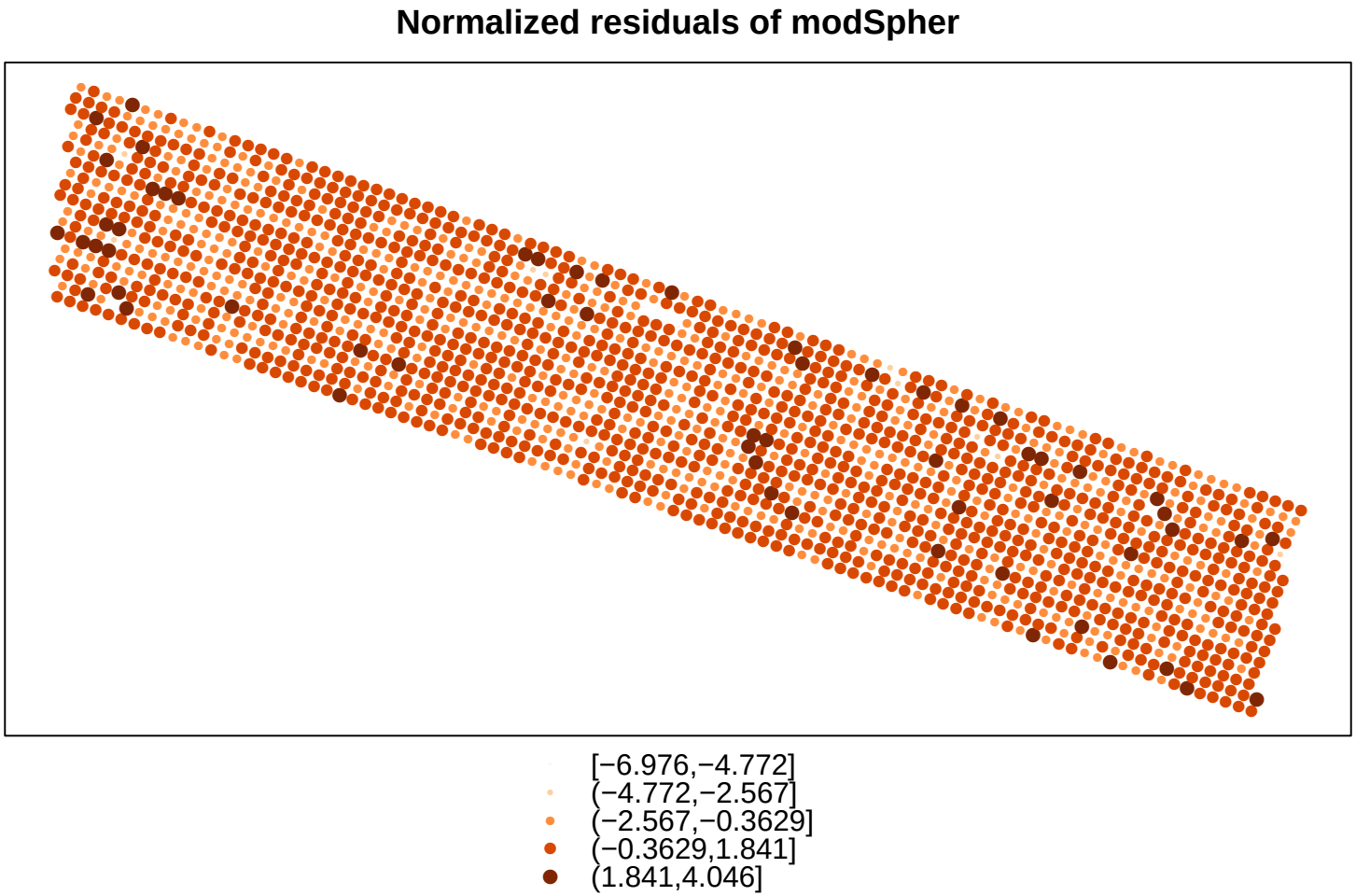


Figure 15 – Bubble map for residuals of `modSpher`.

Example Las Rosas (6/7)

```
moran.mc(Xutm@data$resSpherNorm,W,nsim=1000,alternative="greater")

##
##  Monte-Carlo simulation of Moran I
##
## data:  Xutm@data$resSpherNorm
## weights: W
## number of simulations + 1: 1001
##
## statistic = -0.026172, observed rank = 14, p-value = 0.986
## alternative hypothesis: greater
```

Example Las Rosas (7/7)

```
ks.test(Xutm@data$resSpherNorm,"pnorm", mean=0, sd=sd(Xutm@data$resSpherNor

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  Xutm@data$resSpherNorm
## D = 0.03055, p-value = 0.08311
## alternative hypothesis: two-sided
```

Example Las Rosas (8/7)

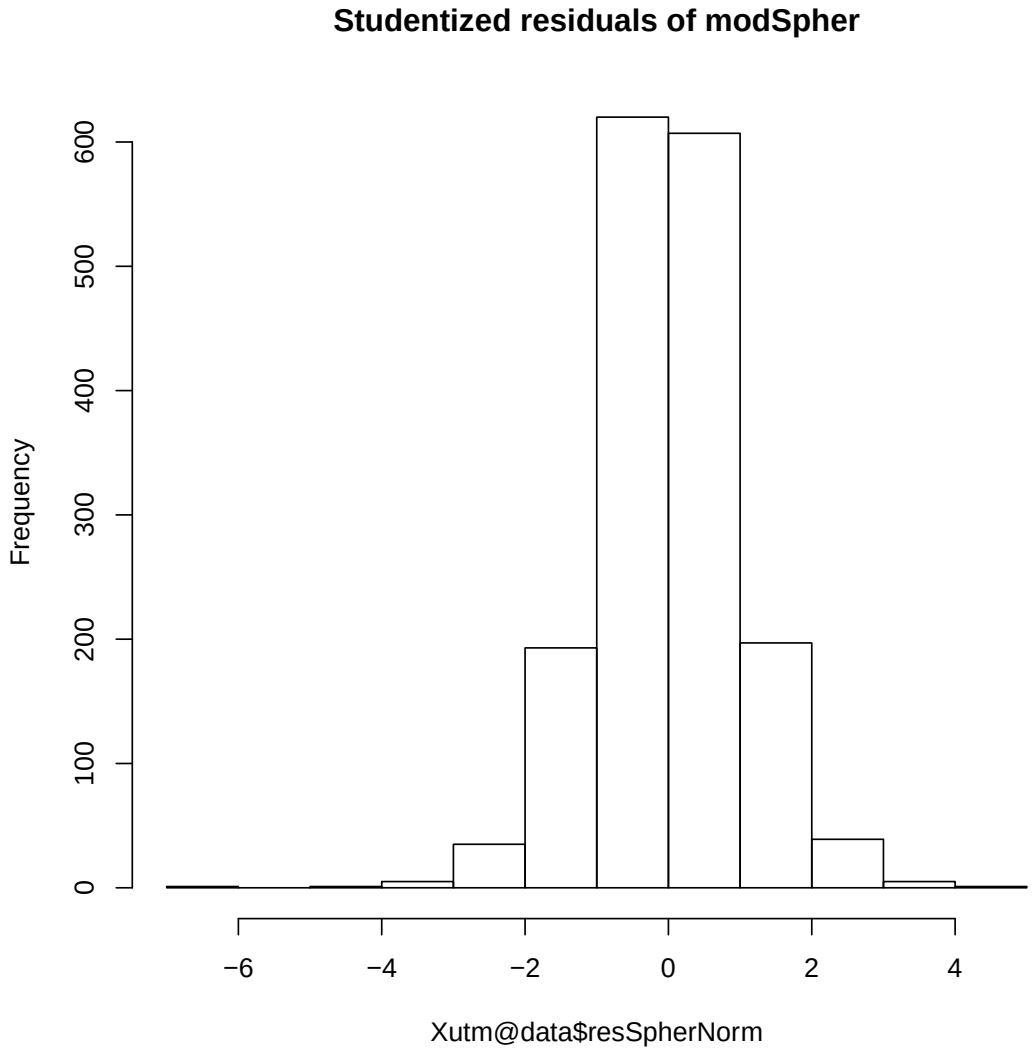


Figure 16 – Histogram of residuals of `modSpher`.

Remark

In practice

- Easier to use a regression model designed for spatially autocorrelated data, and often more intuitive.

- If one of these two models does not give a satisfactory result, you can try an extended linear model $\Rightarrow$ choose the form of Λ yourself, using the form of the semi-variogram or criteria like AIC.

Table of Contents

1 Sources and Consequences of Spatial Autocorrelation

2 Working example : Las Rosas

3 Spatial Lag Model

4 Spatial Error Model

5 Choosing Between Spatial Lag and Spatial Error models

6 Extended Linear Models

7 Bibliography

Bibliography

1. Mixed Effects Models and Extensions in Ecology with R. A. Zuur, E.N. Ieno, N. Walker, A.A. Saveliev and G.M. Smith, Springer, 2009.
2. Spatial Data Analysis in Ecology and Agriculture using R. R. E. Plant, CRC Press, 2012.
3. Mixed-Effects Models in S and S-PLUS. J.C. Pinheiro and D.M. Bates, Springer, 2000.
4. Spatial Processes, Models and Applications. A.D. Cliff and J.K. Ord, Pion Limited, 1981.
5. Statistics for Spatial Data, revised edition. N.A.C. Cressie, Wiley Classics Library, 2015.