Day 3

Correlation Between Two Processes

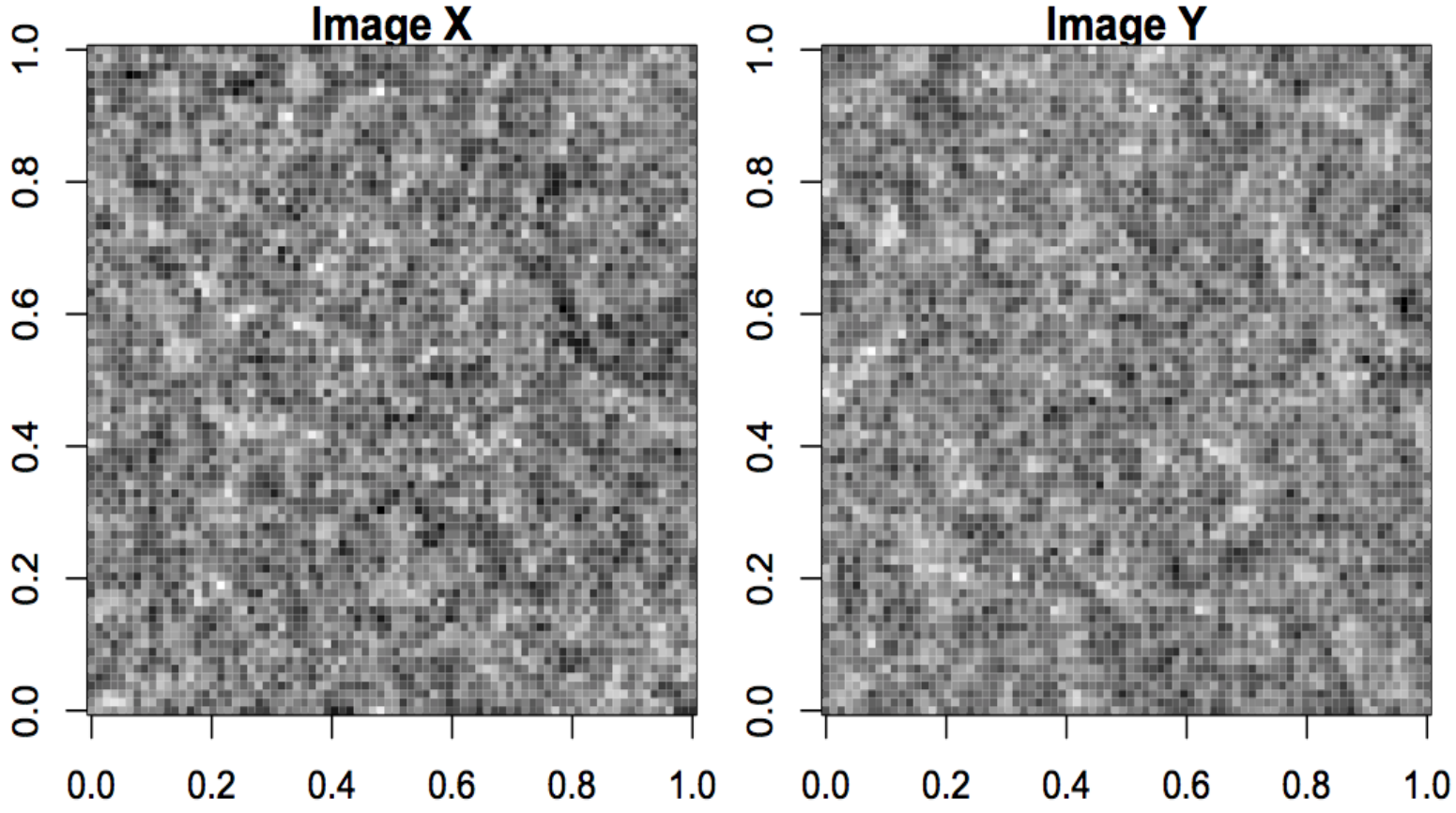
- Let $(X(\boldsymbol{s}), Y(\boldsymbol{s}))^\top$, $\boldsymbol{s} \in \mathbb{R}^2$ be a Gaussian process with mean $(\mu_1, \mu_2)^\top$ and covariance function

$$C(\boldsymbol{h}) = \begin{pmatrix} C_X(\boldsymbol{h}) & C_{XY}(\boldsymbol{h}) \\ C_{YX}(\boldsymbol{h}) & C_Y(\boldsymbol{h}) \end{pmatrix}.$$

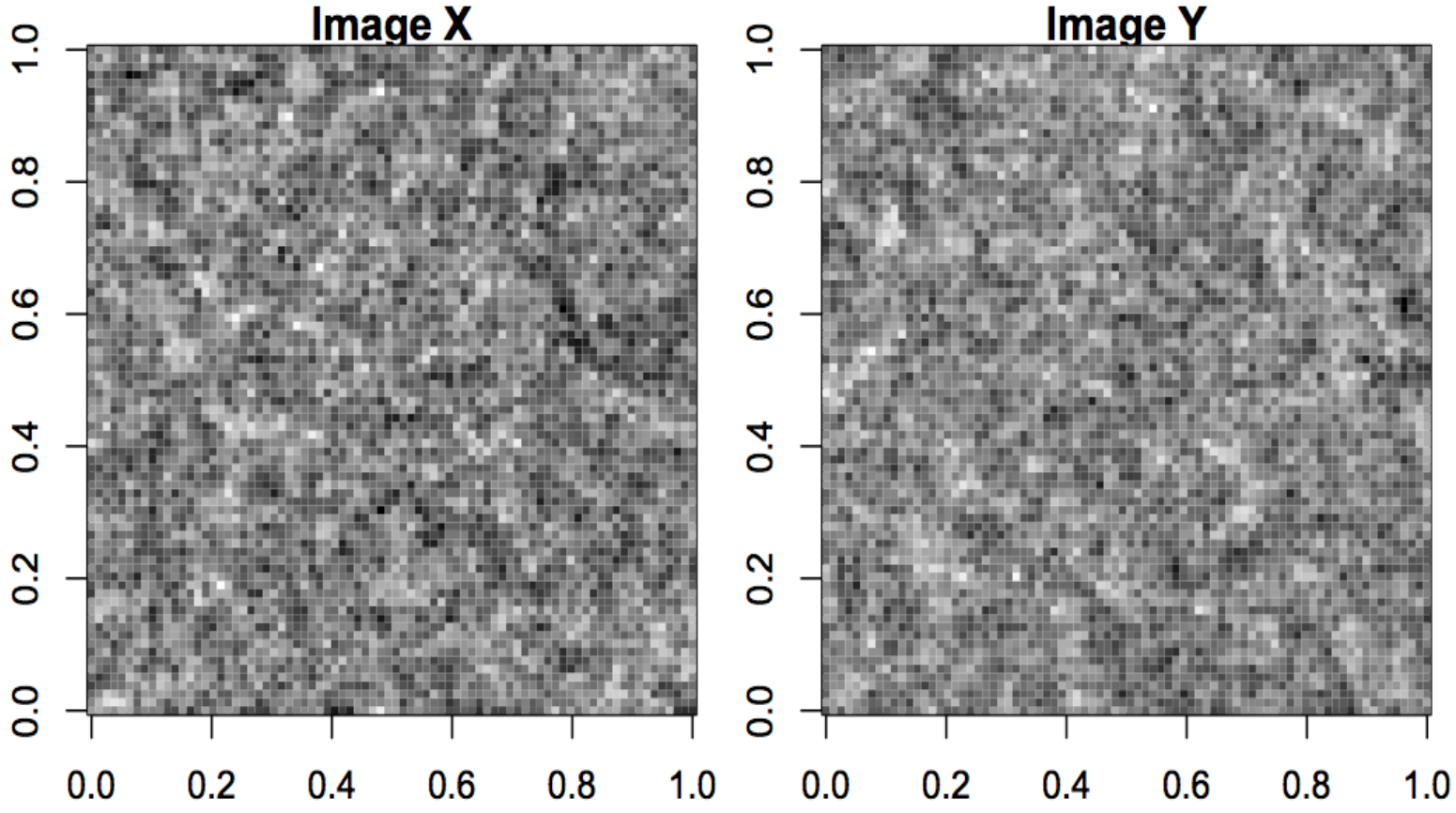
- (Bivarite Matérn)(Gneiting et al. 2010)
  $C_X(\boldsymbol{h}) = \sigma_1^2 M(\boldsymbol{h}, \nu_1, a_1)$,
  $C_Y(\boldsymbol{h}) = \sigma_2^2 M(\boldsymbol{h}, \nu_2, a_2)$, $\mu_1 = \mu_2$,
  $C_{YX}(\boldsymbol{h}, \nu_{12}, a_{12}) = \rho_{12}\sigma_1\sigma_2 M(\boldsymbol{h}, \nu_{12}, a_{12})$,
  where
  $M(\boldsymbol{h}, \nu, a) = (a||\boldsymbol{h}||)^\nu K_\nu(a||\boldsymbol{h}||)$, and $K_\nu(\cdot)$ is a modified Bessel function of the second type and $\rho_{12} = \text{cor}[X(\boldsymbol{s}), Y(\boldsymbol{s})]$.

- How a realization of this process looks like?

Correlation Between Two Processes

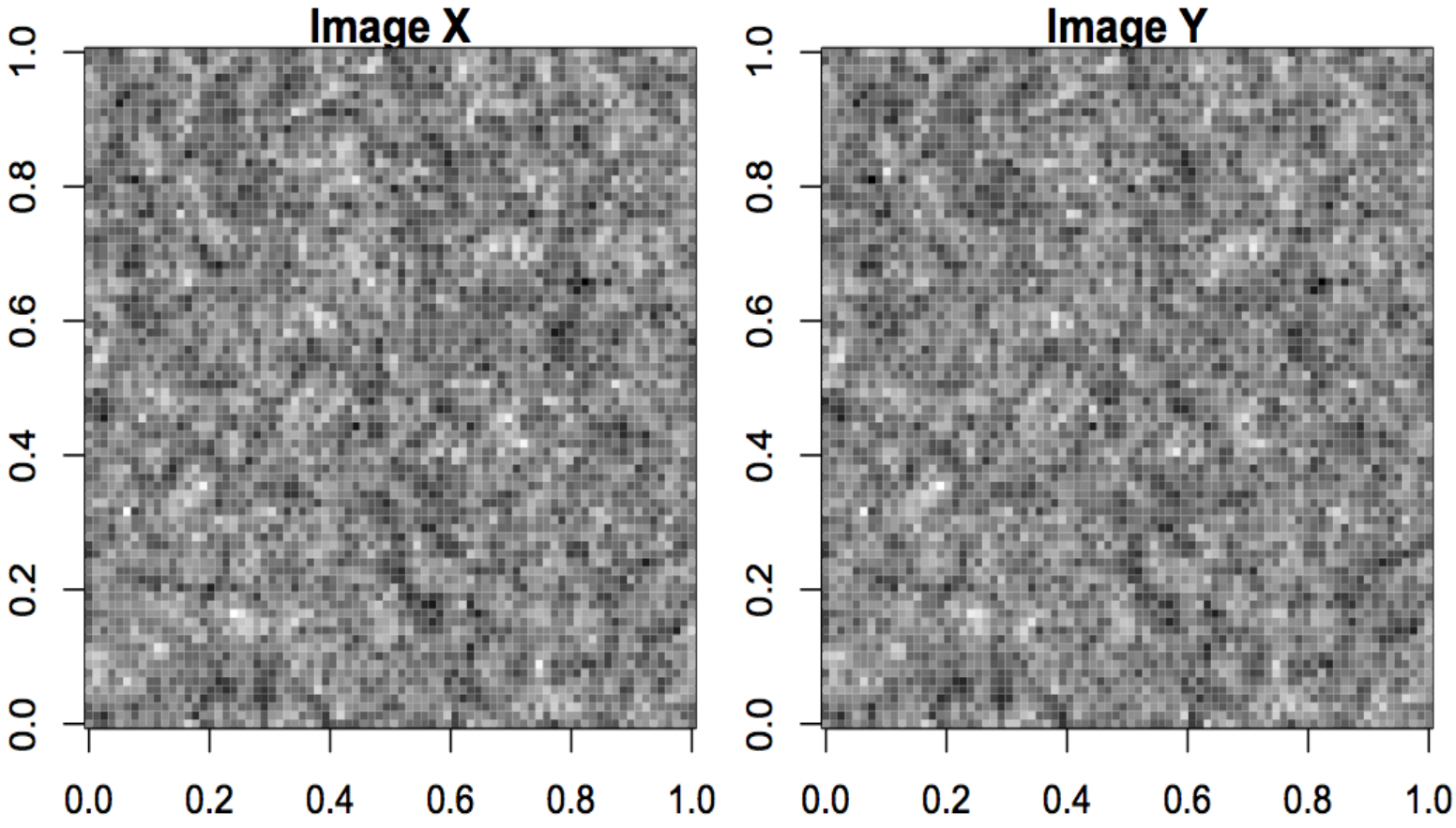


Correlation Between Two Processes

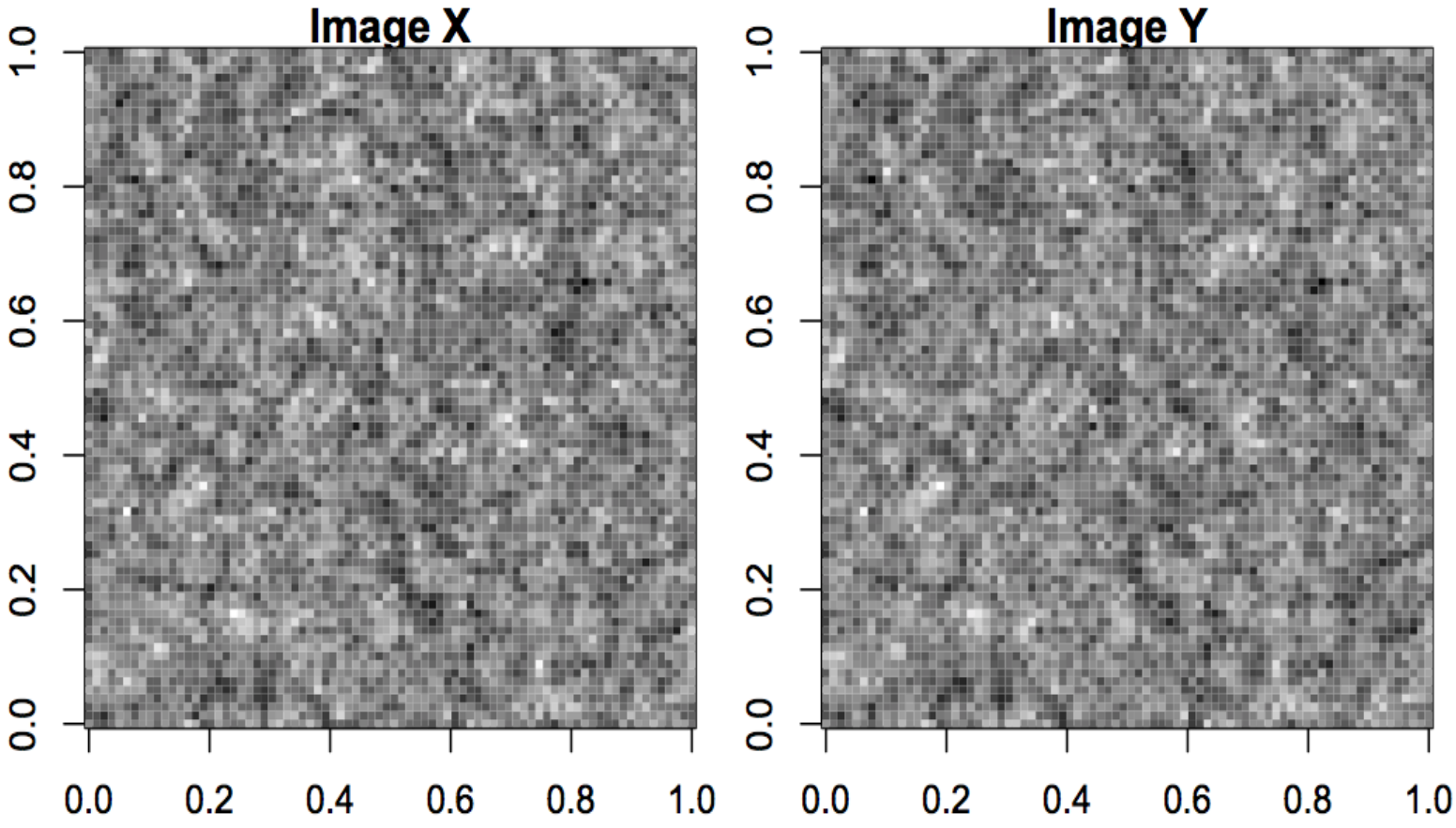


$$\rho_{12} = 0.1$$

Correlation Between Two Processes



Correlation Between Two Processes



$$\rho_{12} = 0.9$$

4. The Modified t Test

- Assume that the two processes $X(\boldsymbol{s})$ and $Y(\boldsymbol{s})$ have been measured on $A = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\} \subset D$.
- Also assume that $\boldsymbol{X} = (X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_n))^\top$ and $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))^\top$ follow a multivariate normal distribution with covariance matrices $\boldsymbol{\Sigma}_X$ y $\boldsymbol{\Sigma}_Y$, respectively.
- A parametric hypothesis testing procedure can be used to dilucidate the hypotheses of presence or absence of correlation between $X(\boldsymbol{s})$ and $Y(\boldsymbol{s})$ considering:

$$\mathrm{H}_0 : \rho_{XY} = 0 \qquad \text{against} \qquad \mathrm{H}_1 : \rho_{XY} \neq 0.$$

- This problem was studied by Cliford et al. (1989) and Dutilleul (1993).

The Modified t Test

- the fundamental idea comes from noticing that the statistic

$$t = \frac{r_{XY}\sqrt{M-2}}{\sqrt{1-r_{XY}^2}} \underset{approx.}{\sim} t_{M-2}.$$

- M is called the *effective sample size* which is defined through

$$M = 1 + (\text{var}[r_{XY}])^{-1},$$

where $r_{XY} = s_{XY}/\sqrt{s_X^2 s_Y^2}$, $\overline{X} = \frac{1}{n}\sum_i X(\boldsymbol{s}_i)$, $s_X^2 = \frac{1}{n}\sum_i (X(\boldsymbol{s}_i) - \overline{X})^2$, $s_{XY} = \frac{1}{n}\sum_i (X(\boldsymbol{s}_i) - \overline{X})(Y(\boldsymbol{s}_i) - \overline{Y})$, and similarly for $\overline{Y}$ and s_Y^2.

Estimation of the Effective Sample Size

- A first estimator is based on Dutilleul (1993) proposal that consists on a first order approximation for the variance of the correlation coefficient

$$\text{var}[r_{XY}] \approx \frac{\text{var}[s_{XY}]}{\mathbb{E}[s_X^2]\mathbb{E}[s_Y^2]},$$

- Assume that the random vector $\boldsymbol{Z} = (\boldsymbol{X}^\top, \boldsymbol{Y}^\top)^\top \sim \mathcal{N}_{2n}(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)$, with

$$\boldsymbol{\mu}_Z = \begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}, \qquad \text{and} \qquad \boldsymbol{\Sigma}_Z = \begin{pmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_Y \end{pmatrix}.$$

Estimation of the Effective Sample Size

- In order to obtain the variance of the bilinear form $s_{XY} = \boldsymbol{X}^\top \boldsymbol{P}\boldsymbol{Y}/n$, where $\boldsymbol{P} = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}_n$, with $\boldsymbol{J}_n = \mathbf{1}_n\mathbf{1}_n^\top$ an $n \times n$ matrix whose elements are all equal to 1, we write

$$s_{XY} = \boldsymbol{Z}^\top \boldsymbol{A}\boldsymbol{Z}/n, \qquad \text{with} \qquad \boldsymbol{A} = \frac{1}{2}\begin{pmatrix} \boldsymbol{P} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{P} \end{pmatrix}.$$

- For multivariate normal random vectors the variance of a quadratic form takes the form

$$\text{var}[\boldsymbol{Z}^\top \boldsymbol{A}\boldsymbol{Z}] = 2\,\text{tr}(\boldsymbol{A}\boldsymbol{\Sigma}_Z\boldsymbol{A}\boldsymbol{\Sigma}_Z) + 4\boldsymbol{\mu}_Z^\top \boldsymbol{A}\boldsymbol{\Sigma}_Z\boldsymbol{A}\boldsymbol{\mu}_Z.$$

- After simple algebra we obtain

$$\text{tr}(\boldsymbol{A}\boldsymbol{\Sigma}_Z\boldsymbol{A}\boldsymbol{\Sigma}_Z) = \frac{1}{2}\left(\text{tr}(\boldsymbol{P}\boldsymbol{\Sigma}_X\boldsymbol{P}\boldsymbol{\Sigma}_Y) + \text{tr}(\boldsymbol{P}\boldsymbol{\Sigma}_{XY}\boldsymbol{P}\boldsymbol{\Sigma}_{XY})\right).$$

- Because of the independence between $X(\boldsymbol{s})$ and $Y(\boldsymbol{s})$, $\boldsymbol{\Sigma}_{XY} = \mathbf{0}$.

Estimation of the Effective Sample Size

- Using standard results for the expected value of a quadratic form one gets
  - $\mathbb{E}[s_X^2] = \mathbb{E}[\boldsymbol{X}^\top \boldsymbol{P}\boldsymbol{X}]/n = \operatorname{tr}(\boldsymbol{P}\boldsymbol{\Sigma}_X)/n.$
  - $\mathbb{E}[s_Y^2] = \operatorname{tr}(\boldsymbol{P}\boldsymbol{\Sigma}_Y)/n.$
- Finally, we introduce the estimator

$$\widehat{\operatorname{var}}[r_{XY}] = \frac{\operatorname{tr}(\boldsymbol{P}\hat{\boldsymbol{\Sigma}}_X \boldsymbol{P}\hat{\boldsymbol{\Sigma}}_Y)}{\operatorname{tr}(\boldsymbol{P}\hat{\boldsymbol{\Sigma}}_X)\operatorname{tr}(\boldsymbol{P}\hat{\boldsymbol{\Sigma}}_Y)}.$$

This implies that

$$\widehat{M} = 1 + (\widehat{\operatorname{var}}[r_{XY}])^{-1}.$$

Hypothesis Testing

- The decision rule rejects the null hypothesis $H_0 : \rho_{XY} = 0$ if

$$\left| \frac{r_{XY}(\widehat{M}-2)^{1/2}}{\sqrt{1-r_{XY}^2}} \right| > t_{1-\alpha/2}(\widehat{M}-2),$$

where $t_{1-\alpha/2}(\widehat{M}-2)$ represents the upper $(1-\alpha/2)100\%$ quantile from the t distribution with $\widehat{M}-2$ degrees of freedom.

- An equivalent modified F-test can be constructed, based on the statistic

$$F = (\widehat{M}-2)\frac{r_{XY}^2}{1-r_{XY}^2},$$

which under H_0, follows an F distribution with 1 and $\widehat{M}-2$ df. Thus, the null hypothesis $H_0 : \rho_{XY} = 0$ is rejected with a level size α if

$$F > F_{1-\alpha}(1, \widehat{M}-2),$$

Estimation of Σ_X and Σ_Y

- We used Moran's I (Moran, 1950) to obtain estimations for $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$.

$$I_X(k) = \frac{1}{n_k s_X^2} \sum_{s_i, s_j \in D_k} w_{ij}(X(\boldsymbol{s}_i) - \overline{X})(X(\boldsymbol{s}_j) - \overline{X}),$$

  where $n_k = \sum_{s_i, s_j \in D_k} w_{ij}$ with $w_{ij} = 1$ if $\boldsymbol{s}_i, \boldsymbol{s}_j \in D_k$ and $w_{ij} = 0$, otherwise. The definition of $I_Y(k)$ is analogous.

- We recall that n_k corresponds to the cardinality of the set D_k.
- Clifford and Richardson (1985) used the estimator

$$\widehat{C}_X(k) = \sum_{s_i, s_j \in D_k} (X(\boldsymbol{s}_i) - \overline{X})(X(\boldsymbol{s}_j) - \overline{X})/n_k.$$

  which led to an equivalent version of the modified t statistic

$$W = n\, s_{XY} \left(\sum_k n_k \widehat{C}_X(k) \widehat{C}_Y(k) \right)^{-1/2}.$$

- $W = (\widehat{M} - 1)^{-1/2} r_{XY}$.

Computational Implementation of the Test

- The modified t-test has been computationally implemented in SpatialPack throughout the function `modified.ttest`.
- The test uses $F = (\widehat{M} - 2) r_{XY}^2 / (1 - r_{XY}^2)$, which (under the null hypothesis) follows an F distribution with 1 and $\widehat{M} - 2$ degrees of freedom.
- The C code underlying the function `modified.ttest` computes an estimation of the effective sample size $\widehat{M} = 1 + (\widehat{\text{var}}[r_{XY}])^{-1}$ where $\widehat{\text{var}}[r_{XY}]$ is computed using Moran's index.
- The output of the function `modified.ttest` is an object of `mod.ttest` class, which includes
  - The F statistic (`Fstat`).
  - The estimated degrees of freedom (`dof`).
  - The p-value associated to the test (`p.value`).
  - The upper bounds for the classes (`upper.bounds`).
  - The number of observations in each class (cardinality) (`card`).
  - A $K \times 2$ matrix containing the Moran indices (`imoran`) for each variable under study (where K is the number of classes).

Application 1: Murray Smelter site revisited

- Previously we introduced the Murray smelter site dataset, which consist of 253 georeferenced soil samples of arsenic (As) and lead (Pb).
- The following piece of R code allows one to make a graph where the large concentrations of As and Pb are represented with larger diameter bubbles.

```
> library(SpatialPack)
> data(murray)
> xpos <- murray$xpos
> ypos <- murray$ypos
> rad.As <- sqrt(murray$As / pi)
> symbols(xpos, ypos, circles = rad.As, inches = 0.35, + bg = "gray",
xlab = "", ylab = "")
> rad.Pb <- sqrt(murray$Pb / pi)
> symbols(xpos, ypos, circles = rad.Pb, inches = 0.35, + bg = "gray",
xlab = "", ylab = "")
```

Application 1: Murray Smelter site

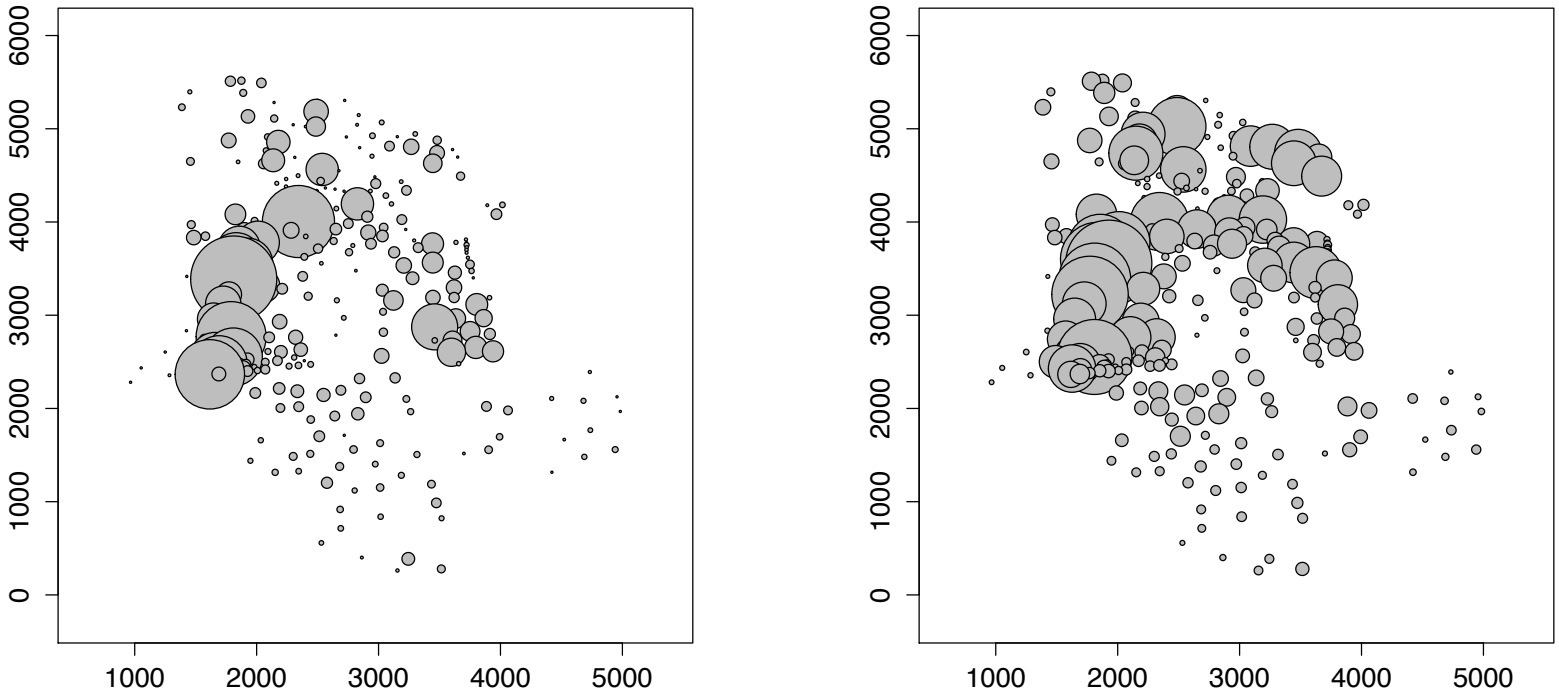


Figure: Bubble plots for (left) As and (right) Pb concentrations.

Application 1: Murray Smelter site

The followiing Rcode runs the Modified t test for the Murray smelter site dataset

```
> coords <- murray[c("xpos", "ypos")]
> x <- murray$As
> y <- murray$Pb
> murray.test <- modified.ttest(x, y, coords)
> class(murray.test)
[1] "mod.ttest"
```

The Routput is

```
> murray.test # output for the modified t-test
Corrected Pearson's correlation for spatial autocorrelation
data: x and y ; coordinates: xpos and ypos F-statistic: 81.949 on 1 and
154.0617 DF, p-value: 0 alternative hypothesis: true autocorrelation is
not equal to 0 sample correlation: 0.5893
```

Application 1: Murray Smelter site

```
> summary(murray.test)
Corrected Pearson's correlation for spatial autocorrelation
data: x and y ; coordinates: xpos and ypos
F-statistic: 81.949 on 1 and 154.0617 DF, p-value: 0
alternative hypothesis: true autocorrelation is not equal to 0
sample correlation: 0.5893
Upper Bounds Cardinality Moran:x Moran:y
1    424.4        1625     0.168696   0.190601
2    848.8        3590     0.056656   0.042300
3   1273.2        4651      0.003615   0.001929
4   1697.6        5061    -0.025936  -0.008734
5   2122.0        5181    -0.035670  -0.035700
6   2546.4        4274    -0.047373  -0.025357
7   2970.8        3285    -0.036903  -0.031196
8   3395.2        1925    -0.038192  -0.063536
9   3819.6        1194     0.019267  -0.061597
10  4244.0         635     0.059552   0.032902
11  4668.4         300     0.066867   0.013618
12  5092.8         129     0.073100   0.080474
13  5517.2          28     0.076299   0.123326
```

Application 2: Modified t-test between images

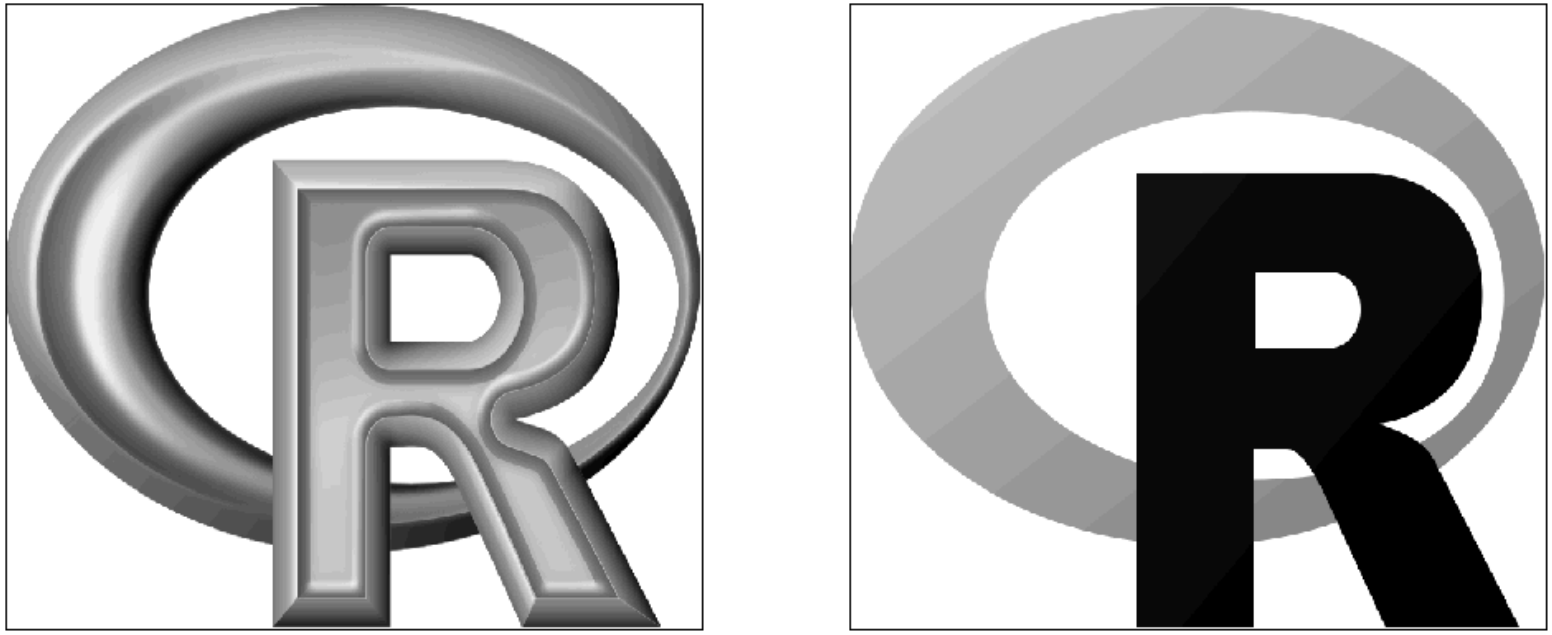


Figure: First band for (Left) old and (Right) new R logos. Both images are of size 561×724. In this case $n = 406\,164$.

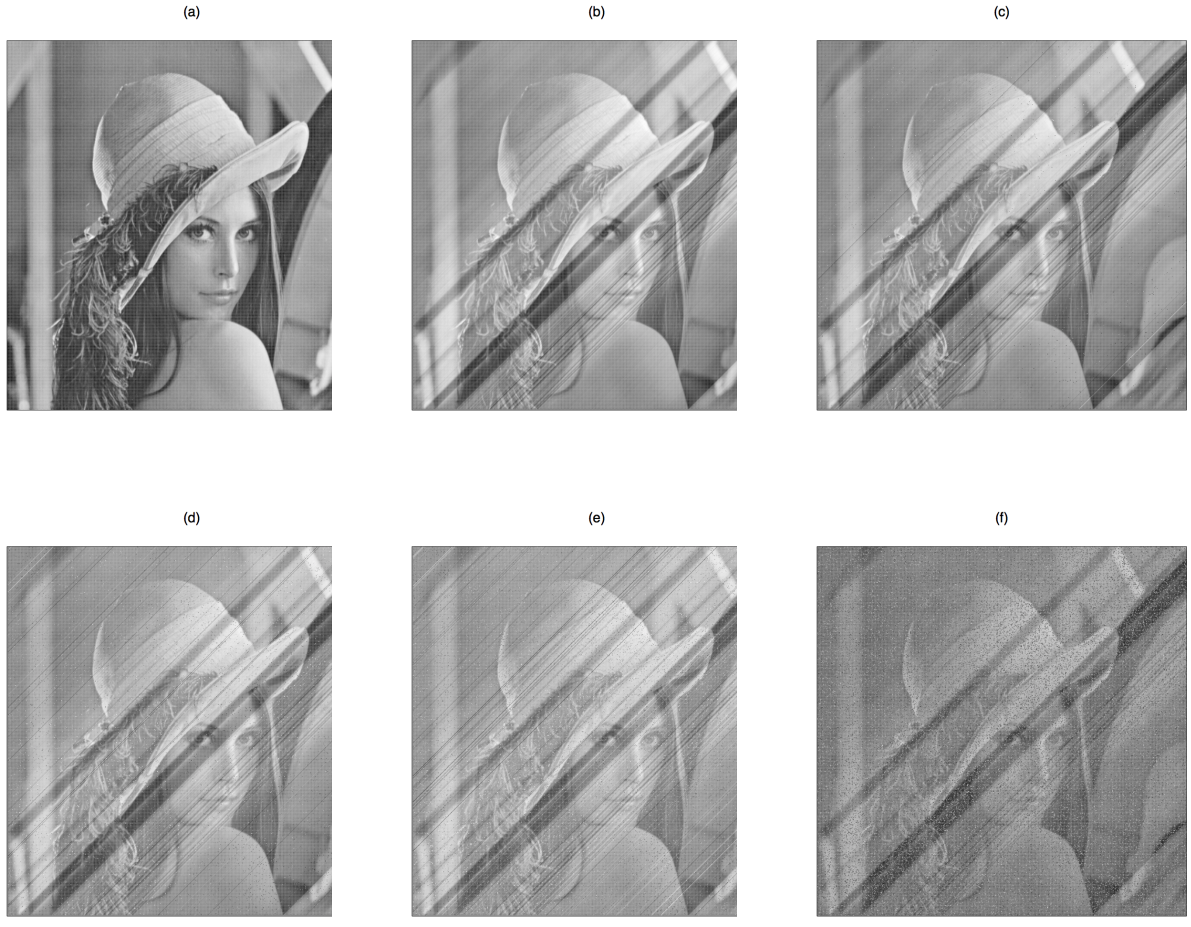
Application 2: Modified t-test between images

```
# coercing images into vectors
> x <- as.vector(oldRlogo)
> y <- as.vector(newRlogo)
# This is a lengthy operation...
> logo.test <- modified.ttest(x, y, coords)
# Output
> logo.test
Corrected Pearson's correlation for spatial autocorrelation
data: x and y ; coordinates: xpos and ypos
F-statistic: 46.5685 on 1 and 117.0716 DF, p-value: 0
alternative hypothesis: true autocorrelation is not equal to 0
sample correlation: 0.5335
```

Application 2: Modified t-test between images

- As expected, the hypothesis of no correlation between the old and new logos should be rejected.
- The sample correlation coefficient is $r_{XY} = 0.5335$, while the effective sample size is $\widehat{M} = 119.0716$.
- $F = 46.5685$ with 1 and 117.0716 degrees of freedom.
- We observe the enormous reduction of the effective sample size in a 99.97%.
- This confirm the strong spatial association between the processes.
- The time required to develop the modified t-test was 5 hours and 30 minutes on a PC with an Intel Core i5 of 2.50 Ghz $\times$ 4 processor and 8 Gb of RAM.

5. Correlation Between One Process and Several Others



Correlation Between One Process and Several Others

- We present an interesting extension of the modified t-test.
- This test takes advantage of the relationship between the correlation coefficient and the determination coefficient in the context of linear regression.
- Dutilleul's methodology can be used for the estimation of the effective sample size.
- Similarly to the modified t this method is for normal random processes.
- This methodology was studied bt Dutilleul et al. (2008).

Correlation Between One Process and Several Others

- Suppose that $\boldsymbol{Z}(\boldsymbol{s}) = (Y(\boldsymbol{s}), \boldsymbol{X}^\top(\boldsymbol{s}))^\top$, is a multivariate process, where $\boldsymbol{s} \in D$ with $\boldsymbol{X}(\boldsymbol{s}) = (X_1(\boldsymbol{s}), \ldots, X_q(\boldsymbol{s}))^\top$.
- We assume that $\boldsymbol{Z}(\boldsymbol{s})$ is a multivariate stationary Gaussian process of size $(q+1)$.
- Based on the set of observations $(Y(\boldsymbol{s}_1), \boldsymbol{X}^\top(\boldsymbol{s}_1))^\top, \ldots, (Y(\boldsymbol{s}_n), \boldsymbol{X}^\top(\boldsymbol{s}_n))^\top$ the goal is to test the hypotheses

$$H_0 : \rho_{Y\boldsymbol{X}} = 0 \qquad \text{against} \qquad H_1 : \rho_{Y\boldsymbol{X}} \neq 0,$$

  where $\rho_{Y\boldsymbol{X}}$ represents the multiple correlation coefficient (Anderson, 2003, pp. 38).
- The sample version of $\rho_{Y\boldsymbol{X}}$ is

$$R_{Y\boldsymbol{X}} = \frac{\boldsymbol{s}_{Y\boldsymbol{X}}^\top \boldsymbol{S}_{\boldsymbol{X}\boldsymbol{X}}^{-1} \boldsymbol{s}_{Y\boldsymbol{X}}}{s_Y^2}.$$

Correlation Between One Process and Several Others

- Anderson (2003) proved that

$$R^2_{Y\boldsymbol{X}} \sim \text{Beta}(q/2, (n-q-1)/2).$$

- Dutilleul et al. (2008) considered the test statistic

$$F = \frac{R^2_{Y\boldsymbol{X}}/q}{(1-R^2_{Y\boldsymbol{X}})/(M-q-1)},$$

where the effective sample size was computed through

$$M = 1 + q/\,\mathbb{E}[R^2_{Y\boldsymbol{X}}].$$

- They proved that under $\text{H}_0 : \rho_{Y\boldsymbol{X}} = 0$,

$$F = \frac{R^2_{Y\boldsymbol{X}}/q}{(1-R^2_{Y\boldsymbol{X}})/(M-q-1)} \underset{approx.}{\sim} F_{(q,M-q-1)}.$$

Correlation Between One Process and Several Others

- $\widehat{M} = 1 + 1/\widehat{\mathbb{E}}[r^2_{Y\widehat{Y}}] = 1 + \dfrac{\mathrm{tr}(\boldsymbol{P}\widehat{\boldsymbol{\Sigma}}_Y)\,\mathrm{tr}(\boldsymbol{P}\widehat{\boldsymbol{\Sigma}}_{\widehat{Y}})}{\mathrm{tr}(\boldsymbol{P}\widehat{\boldsymbol{\Sigma}}_Y\boldsymbol{P}\widehat{\boldsymbol{\Sigma}}_{\widehat{Y}})}$, where $\boldsymbol{\Sigma}_Y$ and $\boldsymbol{\Sigma}_{\widehat{Y}}$ can be estimated as before.
- The critical region associated with the modified F-test is

$$\Big(\frac{\widehat{M}-q-1}{q}\Big)\frac{R^2_{Y\boldsymbol{X}}}{1-R^2_{Y\boldsymbol{X}}} > F_{1-\alpha}(q, \widehat{M}-q-1),$$

where $F_{1-\alpha}(q, \widehat{M}-q-1)$ denotes the upper quantile of order $(1-\alpha)100\%$ from the F distribution with q and $\widehat{M}-q-1$ degrees of freedom.

- The performance of this test was studied by Dutilleul (2008) through a Monte Carlo simulation study.

The Pinus Radiata Dataset Revisited

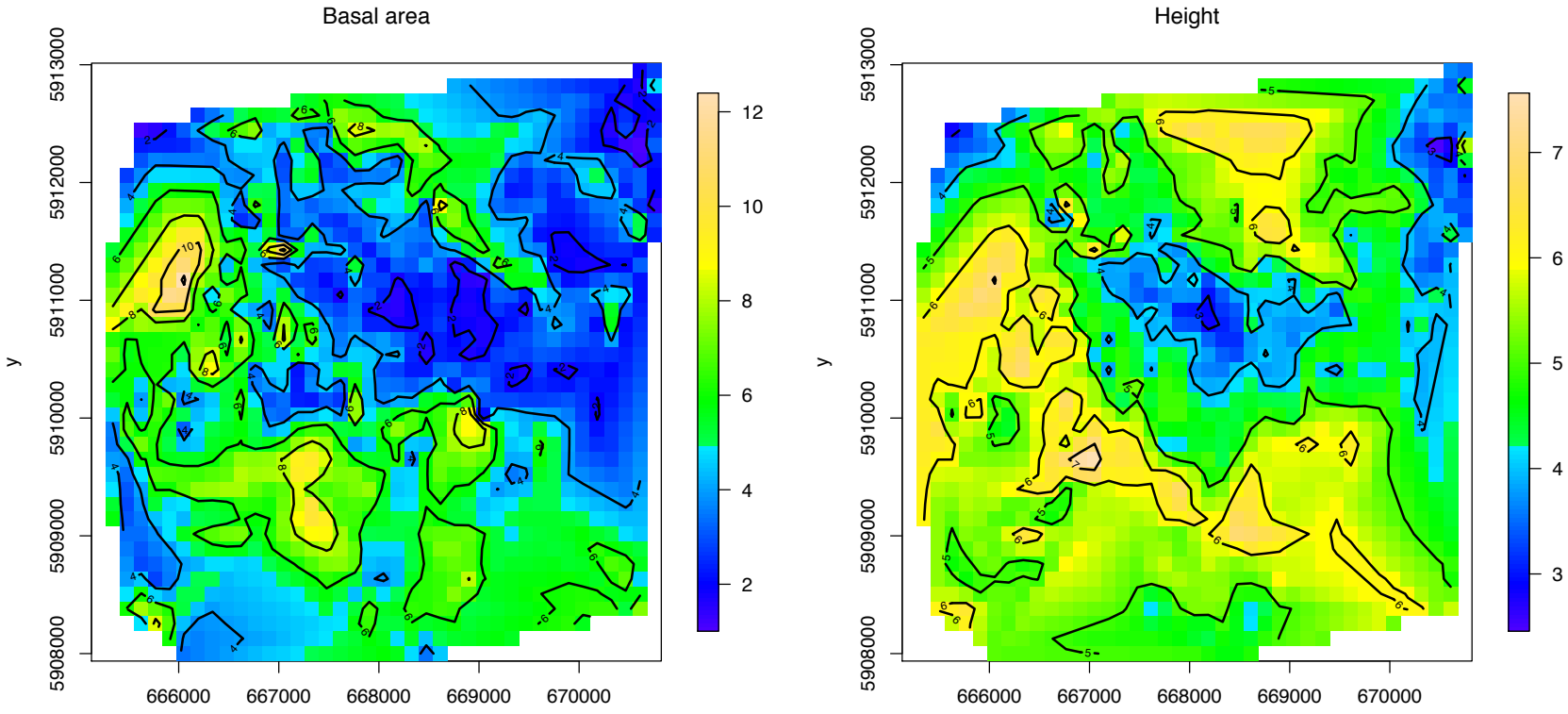


Figure: (Left) Bilinear interpolation of the three basal areas; (Right) Bilinear interpolation of the three heights.

The Pinus Radiata Dataset

- In order to exemplify the methodology, we will assume that the height of the trees (Y) can be influenced by the basal area (X_1), the elevation (X_2) and the slope (X_3) of the terrain.
- The `modified.Ftest` routine is available in the R package `SpatialPack`.
- That routine takes advantage of the C code underlying the routine `modified.ttest`.
- The following code fragment in R presents the results of the modified F-test for the pinus radiata dataset.

The Pinus Radiata Dataset

```
# load the pinus Radiata dataset
> data(radiata)
# defining the response and predictor variables > y <-
radiata$height
> x <- radiata[c("basal","altitude","slope")]
# extracting coordinates
> coords <- radiata[c("xpos","ypos")]
# computing the modified F-test for spatial association
> radiata.test <- modified.Ftest(x, y, coords)
# estimated effective sample size
> radiata.test$ESS
[1] 83.46113
```

The Pinus Radiata Dataset

```
# detailed output for the modified F-test
> summary(radiata.test)
Multiple correlation for assessing spatial autocorrelation
F-statistic: 14.2708 on 3 and 79.4611 DF, p-value: 0
alternative hypothesis: true multiple correlation is not equal to 0
sample correlation: 0.5917
```

As expected, (see also Cuevas et al. 2013), the hypothesis

$$H_0 : \rho_{Y\boldsymbol{X}} = 0$$

is rejected, indicating the presence of spatial association

Day 4