

Estatística Espacial (MI418) / Geoestatística (ME907)

Guilherme Ludwig

2019-01-14

Simulação de Gaussian random fields

Amostragem e variograma empírico

Resolução, *practical range* e estimabilidade

Variogramas são c.n.d.

Eu esqueci de mencionar essa propriedade: variogramas 2γ são **condicionalmente** negativos definidos, isto é: para todo $n \geq 1$, para todo $(a_1, \dots, a_n) \in \mathbb{R}^n$, tal que $\sum_{i=1}^n a_i = 0$, e para todo $\mathbf{s}_1, \dots, \mathbf{s}_n$ em \mathcal{D} , temos

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0.$$

Note que a condição $\sum_{i=1}^n a_i = 0$, não existe na definição de positivo/negativo definidas. Logo, todas as funções p.d. são c.p.d. mas nem todas as funções c.p.d. são p.d.. Por exemplo, a função $f(\mathbf{h}) = \|\mathbf{h}\|^2 \log(\|\mathbf{h}\|)$ é c.p.d. mas não é p.d. (Wahba, 1990).

Também errei: $\gamma(\mathbf{0}) = 0$, nugget é o limite quando $\mathbf{h} \rightarrow \mathbf{0}$.

Exemplo

- ▶ Vou mostrar alguns exemplos de realizações de processos estocásticos, que ajudarão a entender alguns conceitos teóricos da aula passada.
- ▶ Um objetivo secundário é introduzir alguns dos pacotes estatísticos de análise de dados espaciais. Há muitos, com muitas aplicações em comum. Eu vou considerar (mais ou menos na ordem que eles foram introduzidos no R)
 - ▶ `fields` (ferramentas básicas, Kriging),
 - ▶ `geoR` (vários modelos de covariância, estimação por máxima verossimilhança),
 - ▶ `RandomFields` (amostragem eficiente de campos aleatórios, processos multivariados) e
 - ▶ `gstat` (modelos de kriging mais complexos).

Exemplo

```
require(geoR)
res <- 64 # Resolution of the grid
s1 <- seq(0, 1, length.out = res)
s2 <- seq(0, 1, length.out = res)
Locations <- expand.grid(s1 = s1, s2 = s2)
D <- as.matrix(dist(as.matrix(Locations)))
# Modelo de covariância do geoR
Sigma <- cov.spatial(D, cov.model = "gaussian",
                     cov.pars = c(sigma2 = 1, phi = 0.5))
dim(Sigma)
```

```
## [1] 4096 4096
```

Uma nota

Gerar uma observação \mathbf{x} de um vetor Normal multivariado $n \times 1$ com média $\boldsymbol{\mu}$ (conhecida) e variância $\boldsymbol{\Sigma}$ (conhecida) envolve:

1. Gerar Z_1, \dots, Z_n i.i.d. Normal (por exemplo, com `rnorm`).
2. Uma amostra \mathbf{x} é obtida por $\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}$.

Aqui $\boldsymbol{\Sigma}^{1/2}$ é qualquer decomposição tal que $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{t/2} = \boldsymbol{\Sigma}$, por exemplo Choleski, $\mathbf{Q}\boldsymbol{\Lambda}^{1/2}$, etc.

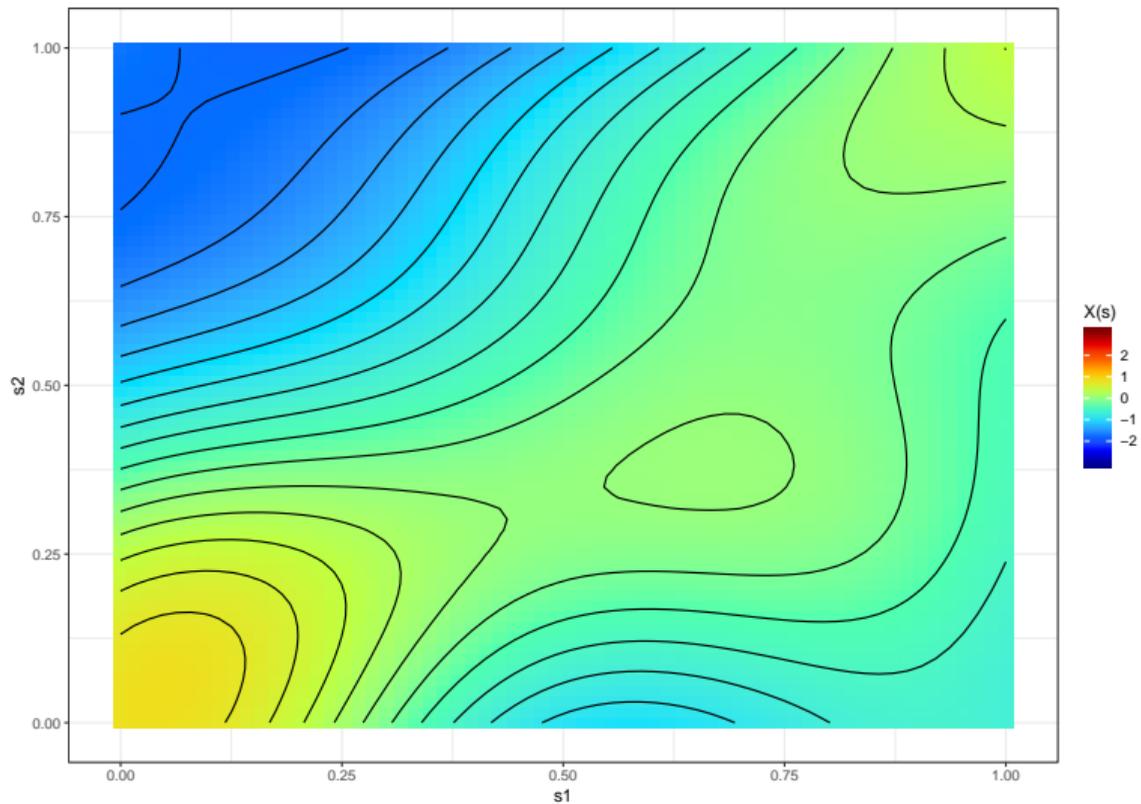
Exemplo

O método mais antigo de geração de Gaussian random fields... é gerar um vetor multivariado! Mas é caro computacionalmente.

```
require(MASS)
set.seed(1)
system.time({
  x <- mvrnorm(1, mu = rep(0, nrow(Locations)),
              Sigma = Sigma) # Inefficient
})
```

```
##      user  system elapsed
## 85.567  12.572  26.671
```

Exemplo



Modelos de variograma

A princípio, vamos considerar apenas processos estacionários. Então

$$2\gamma(\mathbf{h}) = \text{Var}(X(\mathbf{s}) - X(\mathbf{s} + \mathbf{h})) = 2C(\mathbf{0}) - 2C(\mathbf{h}).$$

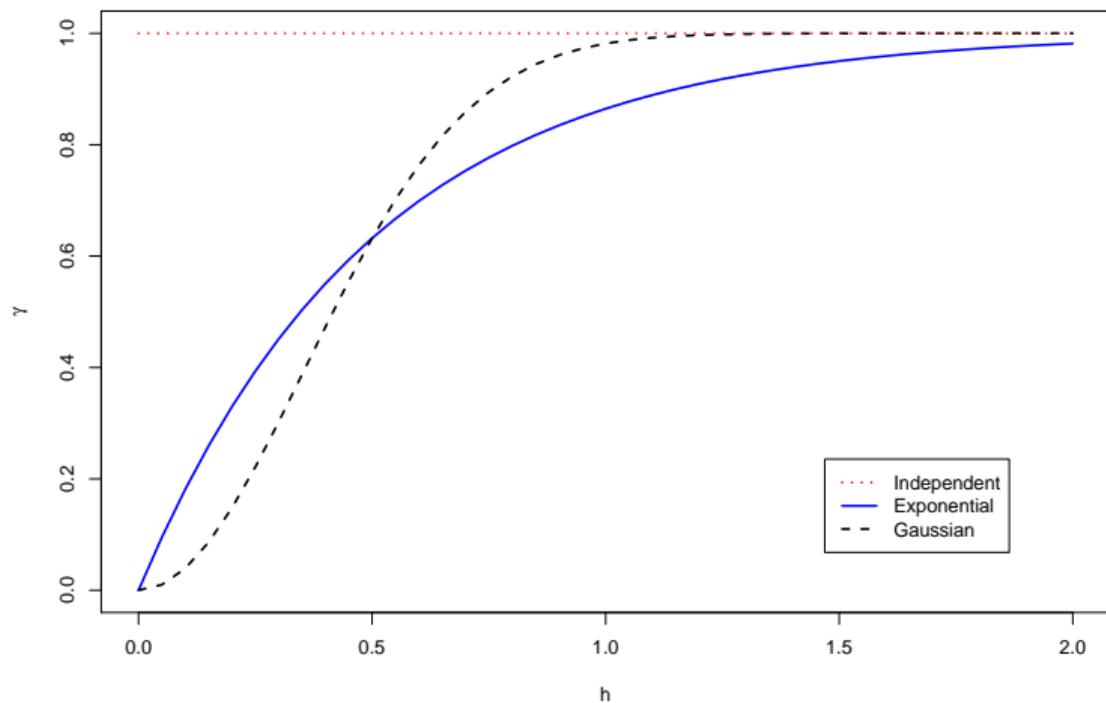
Naturalmente,

$$C(\mathbf{h}) = C(\mathbf{0}) - \gamma(\mathbf{h}).$$

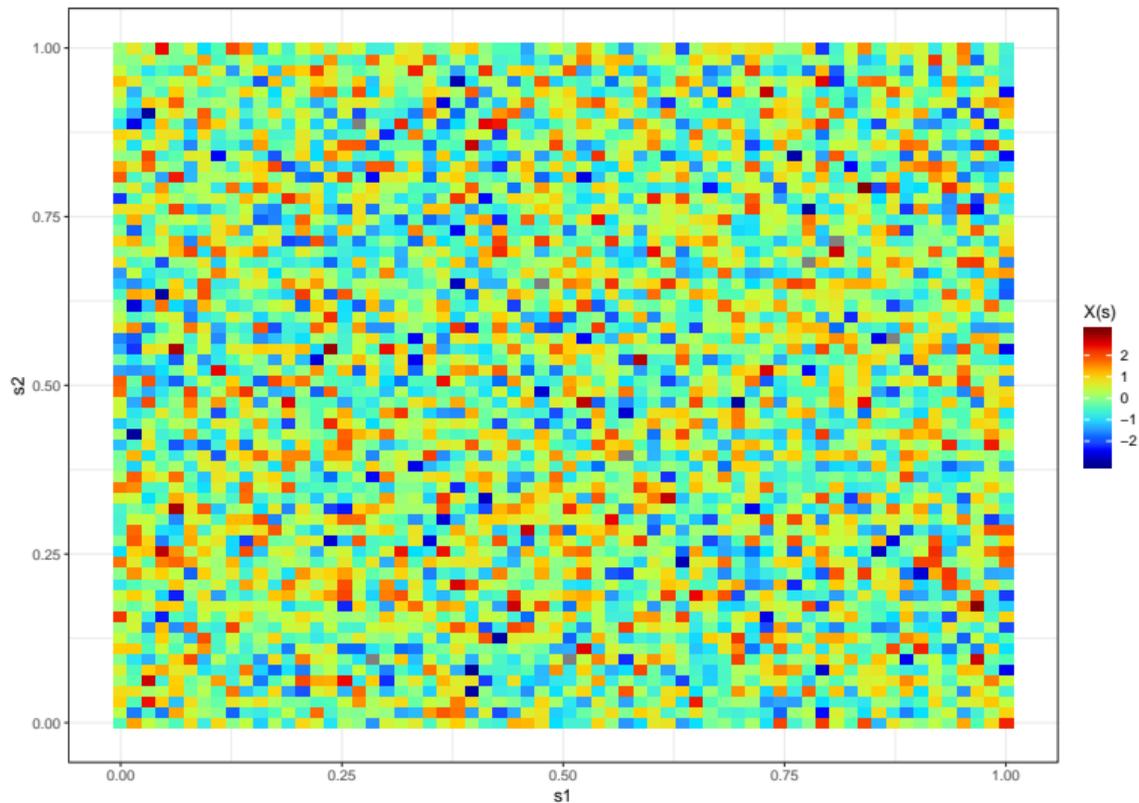
Alguns exemplos de modelos de variograma:

- ▶ *Independente*: $\gamma_0(\mathbf{h}) = \sigma^2$.
- ▶ *Exponencial*: $\gamma_1(\mathbf{h}) = \sigma^2 (1 - \exp\{-\|\mathbf{h}\|/\phi\})$.
- ▶ *Gaussiano*: $\gamma_2(\mathbf{h}) = \sigma^2 (1 - \exp\{-\|\mathbf{h}\|^2/\phi^2\})$.

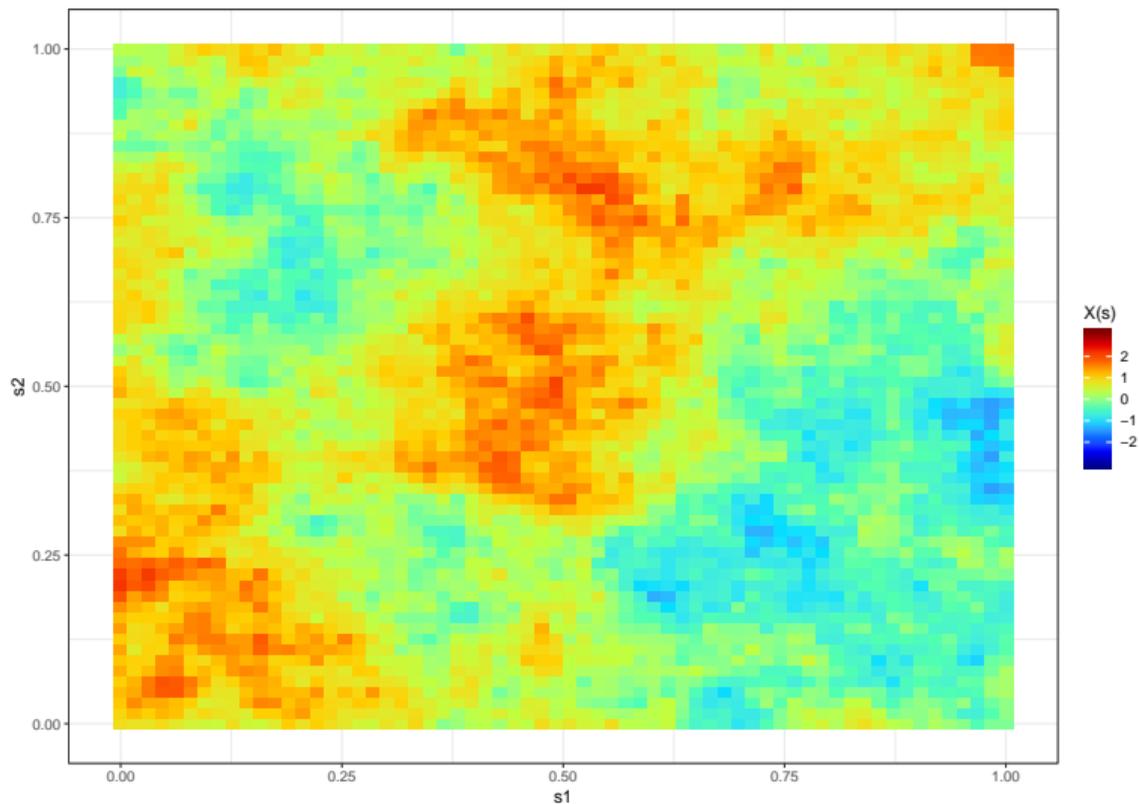
Modelos de variograma ($\sigma^2 = 1, \phi = 0.5$)



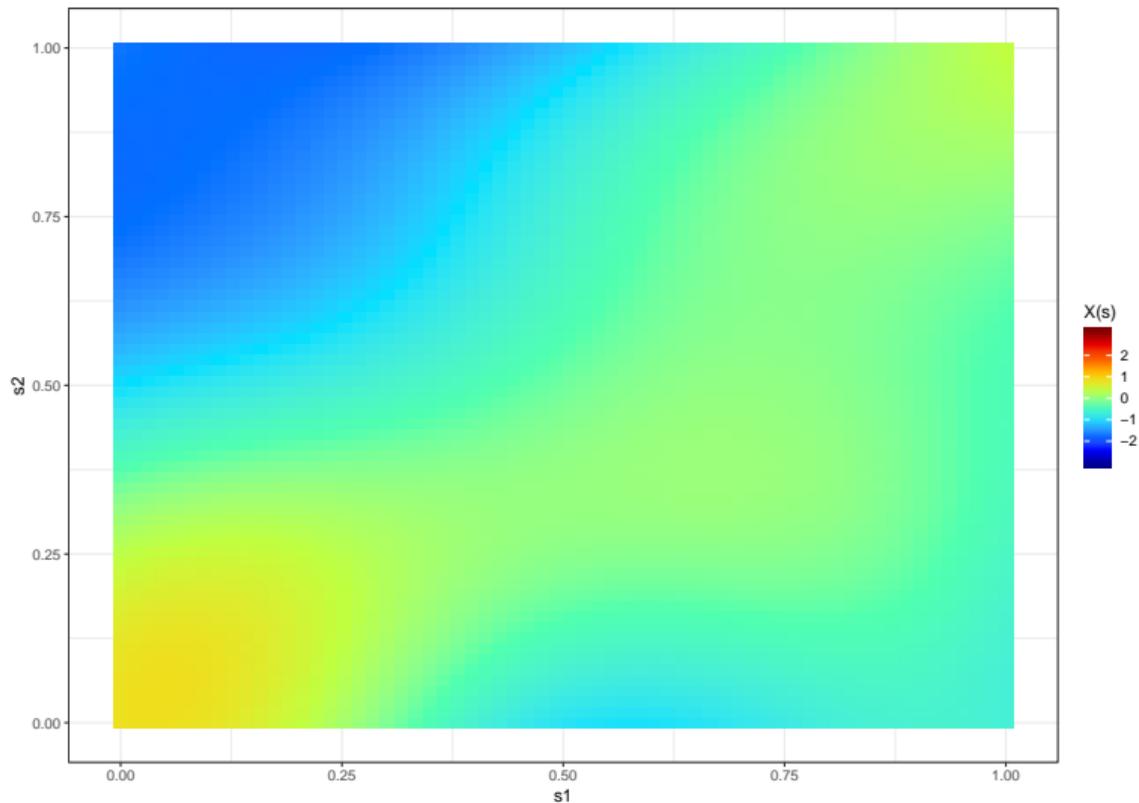
Processo Gaussiano com variograma independente



Processo Gaussiano com variograma exponencial



Processo Gaussiano com variograma Gaussiano



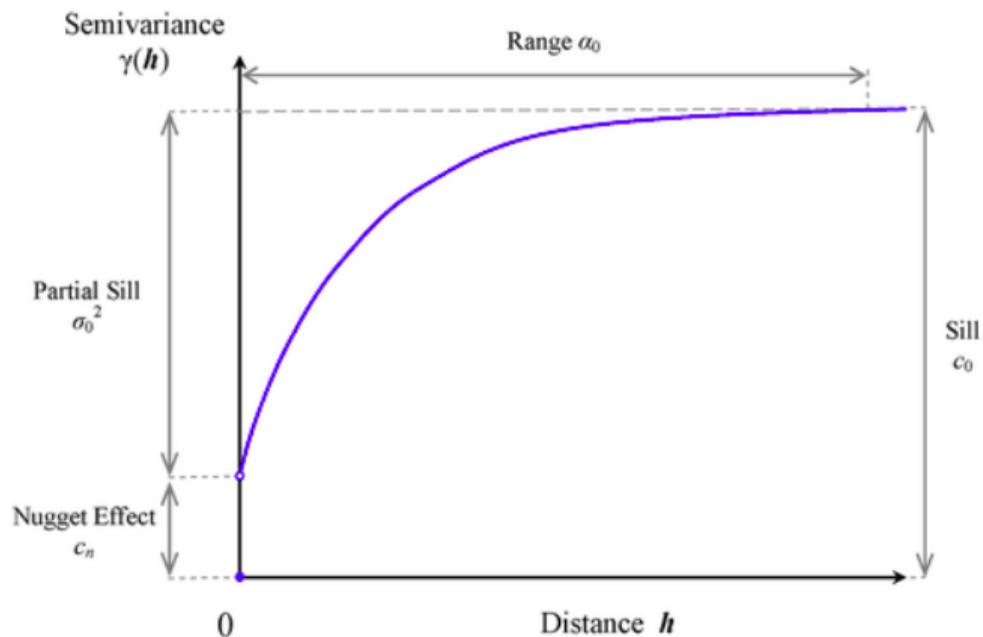
Visualizando propriedades teóricas

- ▶ Um variograma independente não é contínuo em $\mathbf{h} = 0$. Logo, uma observação de um processo Gaussiano com esse variograma não é contínua.
- ▶ Um variograma exponencial é contínuo em $\mathbf{h} = 0$, porém, não é diferenciável (cf. lista #1). Logo, uma observação de um processo Gaussiano com esse variograma é contínua mas não é diferenciável.
- ▶ Um variograma Gaussiano é infinitamente diferenciável em $\mathbf{h} = 0$. Logo, uma observação de um processo Gaussiano com esse variograma é diferenciável.

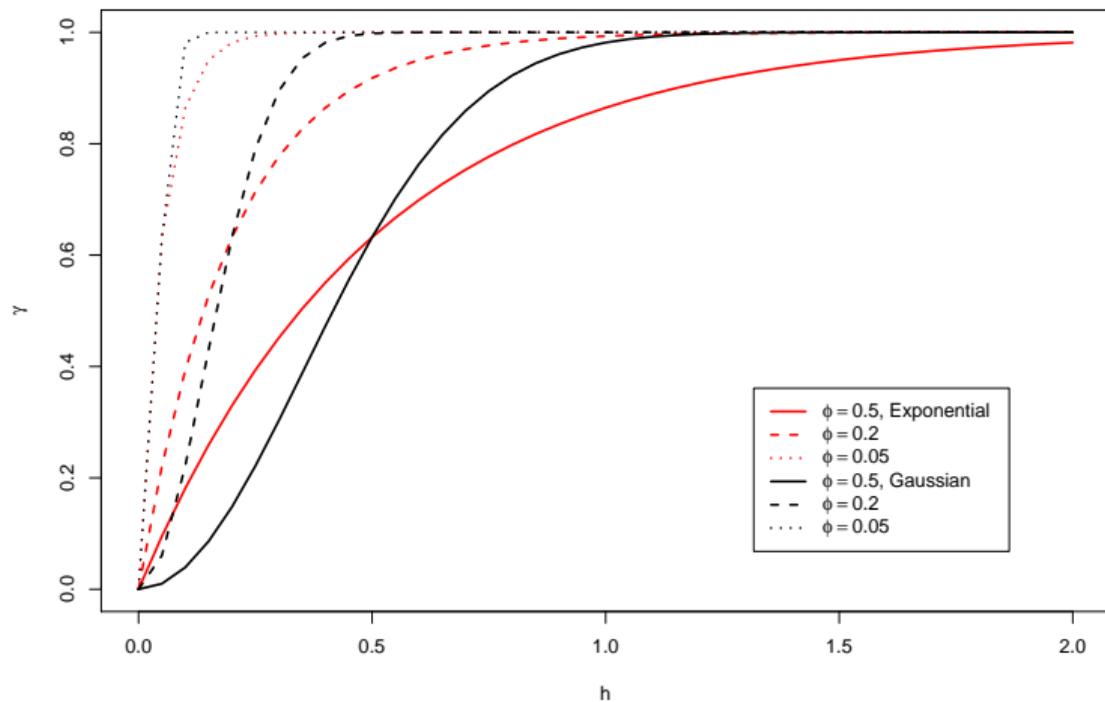
Nunca é demais lembrar: na prática, não observamos o processo em todo o domínio $\mathbf{s} \in \mathcal{D}$, apenas em uma coleção de pontos $\mathbf{s}_1, \dots, \mathbf{s}_n$.

Nomenclatura

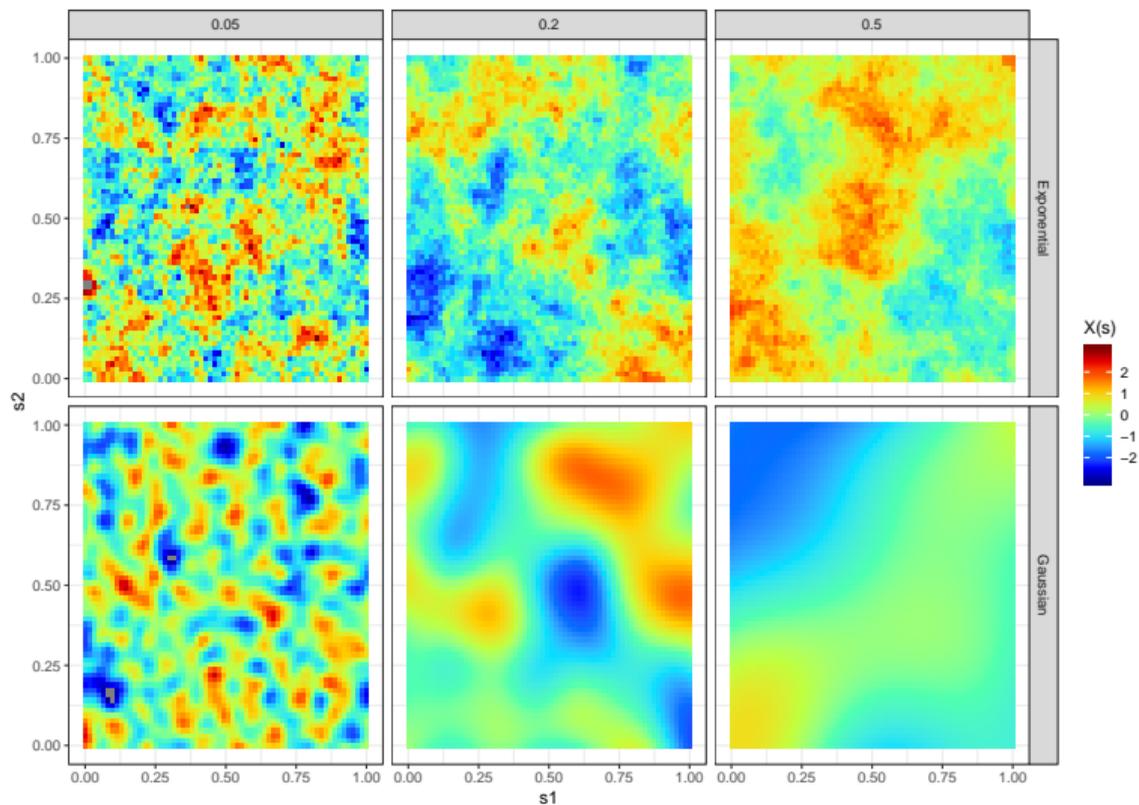
Figura extraída de http://support.sas.com/documentation/cdl/en/statug/66859/HTML/default/statug_variogram_details02.htm



Interesse no range of dependence (ϕ)



Interesse no range of dependence (ϕ)



Efeito de nugget

O nugget é o efeito de erro instrumental no campo Gaussiano. Você sempre pode pensar que o modelo de covariância com nugget vem do processo

$$Y(\mathbf{s}) = Z(\mathbf{s}) + \varepsilon(\mathbf{s}),$$

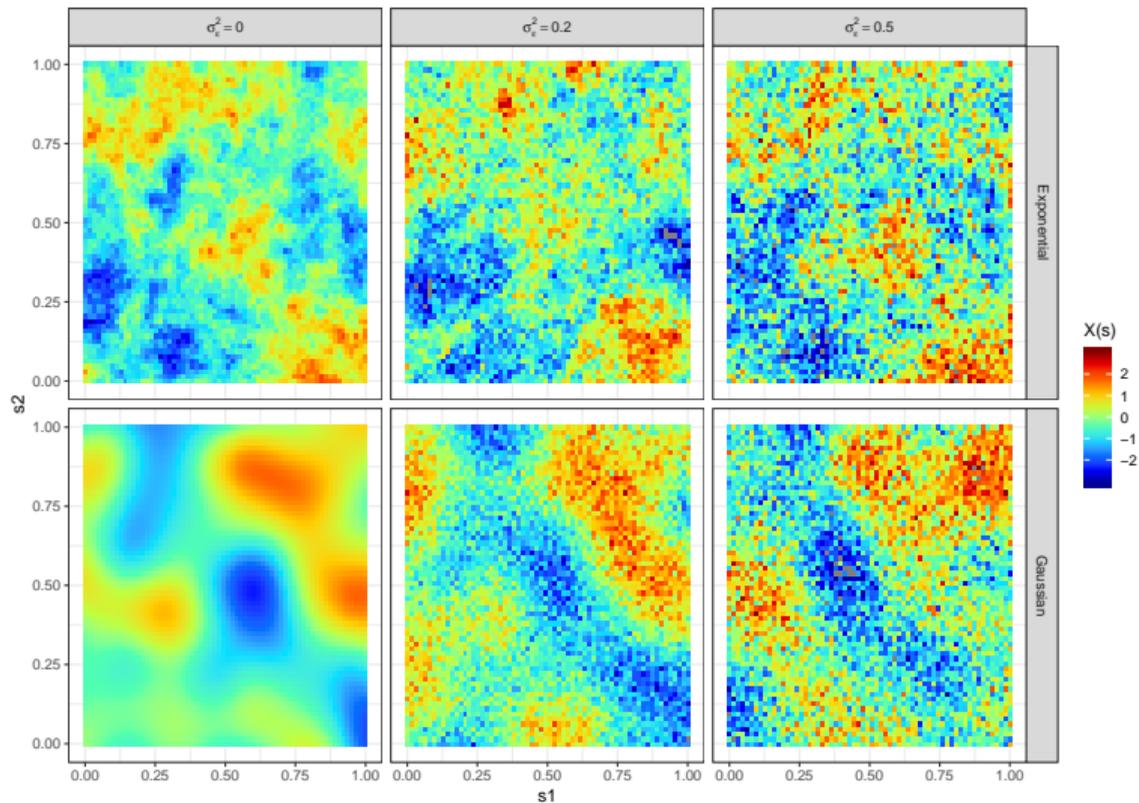
em que $Z(\mathbf{s})$ é um processo Gaussiano (estacionário, isotrópico, ou o que você preferir) e $\varepsilon(\mathbf{s})$ é um processo de erro *white noise*, independente de $Z(\mathbf{s})$.

É fácil ver que se $Z(\mathbf{s})$ tem covariância $C(\mathbf{h}) = \sigma_Z^2 \rho(\|\mathbf{h}\|/\phi)$, então $\text{Cov}(Z(\mathbf{s}) + \varepsilon(\mathbf{s})) = \sigma_\varepsilon^2 + \sigma_Z^2 \rho(\|\mathbf{h}\|/\phi)$. Consequentemente,

$$\gamma(\mathbf{h}) = \sigma_\varepsilon^2 + \sigma_Z^2(1 - \rho(\|\mathbf{h}\|/\phi)).$$

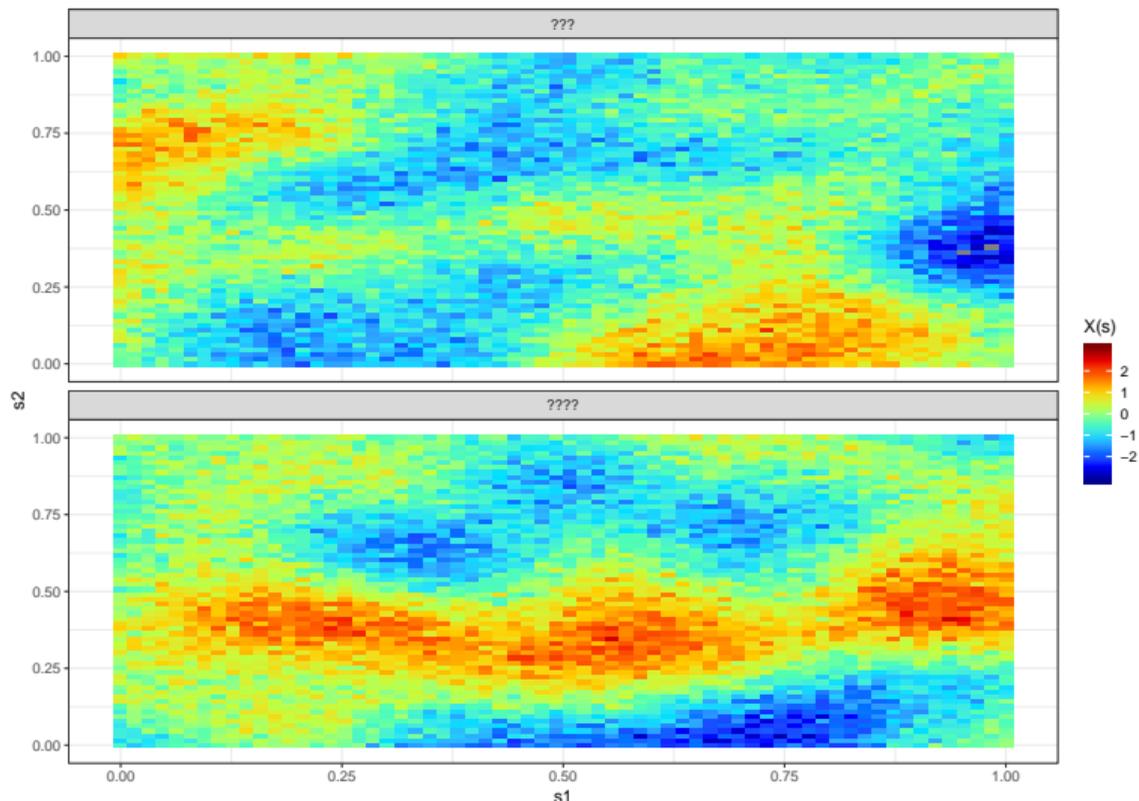
De fato, $\sigma_Z^2/\sigma_\varepsilon^2$ é chamada *signal-to-noise* ratio. O nugget tem outra interpretação importante que consideraremos adiante. . .

Efeito do nugget ($\phi = 0.2, \sigma_Z^2 = 1$)

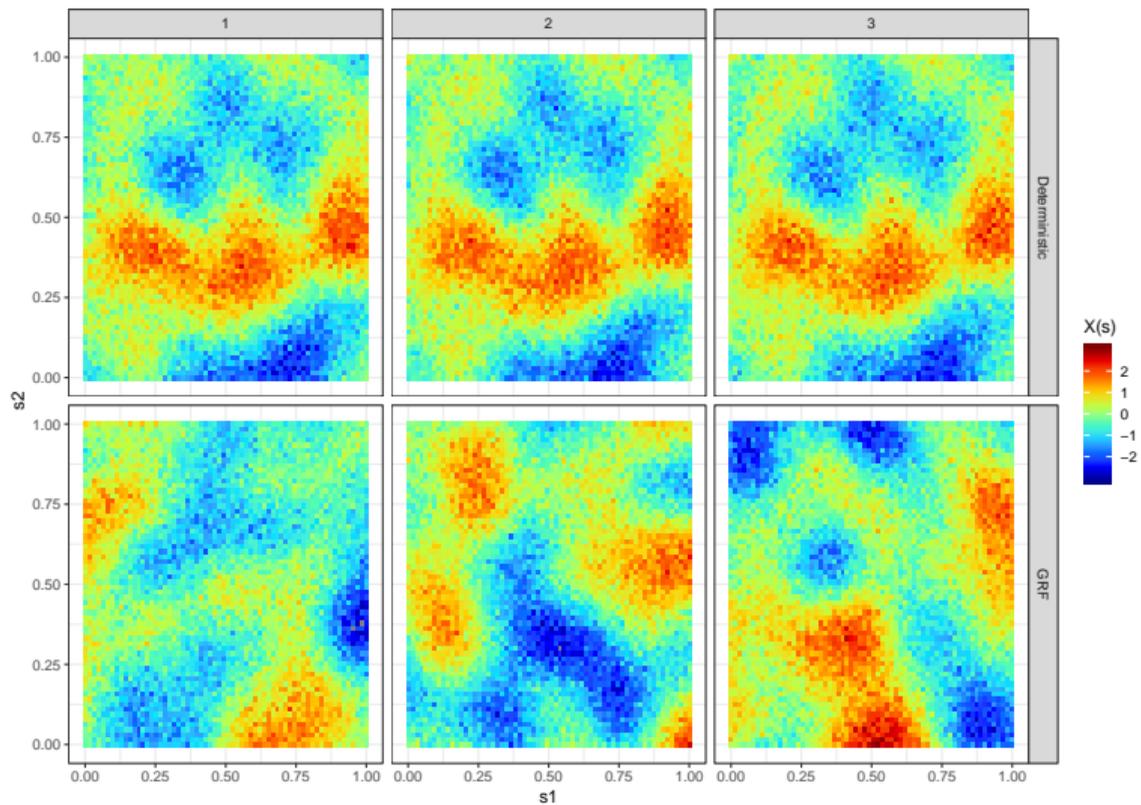


Uma reflexão

A figura a seguir tem dois processos: um Gaussiano, estacionário, + erro, e outro simplesmente uma função determinística + erro. Você consegue distinguir?



Replicações



Não há replicações em estudos espaciais.

Não há replicações em estudos espaciais.¹

¹Exceto em uma fração pequena de casos planejados... ou seja, qualquer metodologia que dependa de replicações é questionável.

Alguns modelos extras

- ▶ Matérn: dado por

$$\rho(\mathbf{h}) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\|\mathbf{h}\|}{\phi} \right)^{\nu} \mathcal{K}_{\nu} \left(\frac{\|\mathbf{h}\|}{\phi} \right),$$

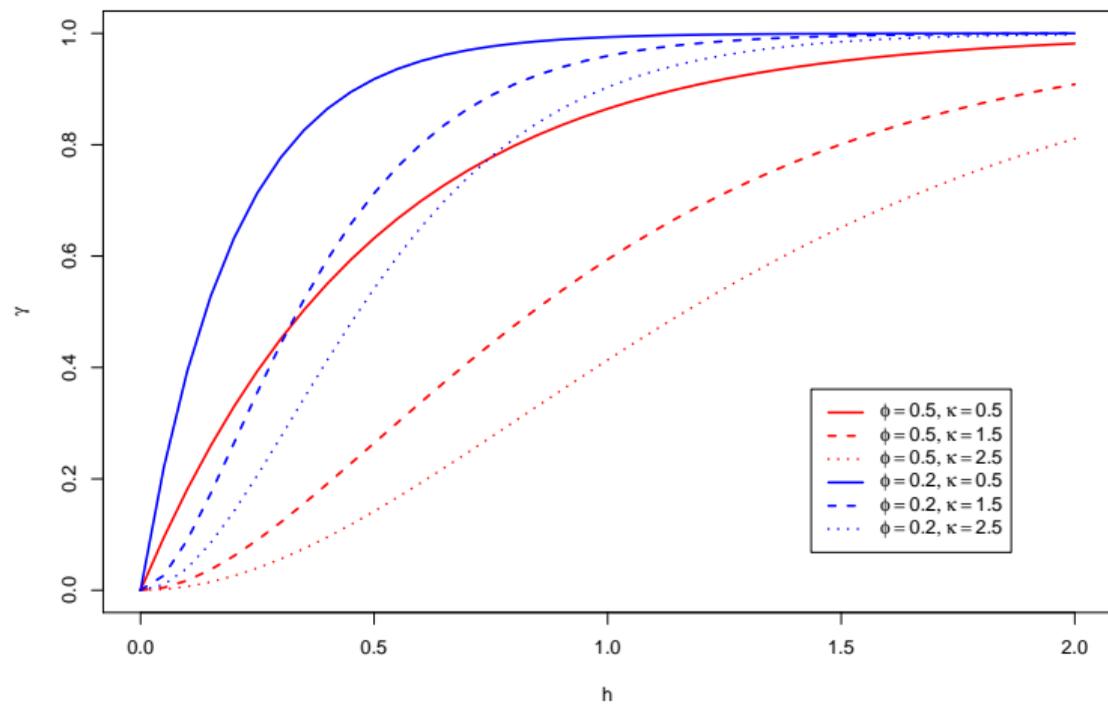
onde \mathcal{K}_{ν} é a função tipo II modificada de Bessel. Esse modelo generaliza o exponencial ($\nu = 1/2$) e Gaussiano ($\nu = \infty$). É particularmente interessante porque, para ν fixo, o processo resultante é ν -vezes L^2 diferenciável. Em geral não se estima ν , cuidado.

- ▶ Esférico: dado por

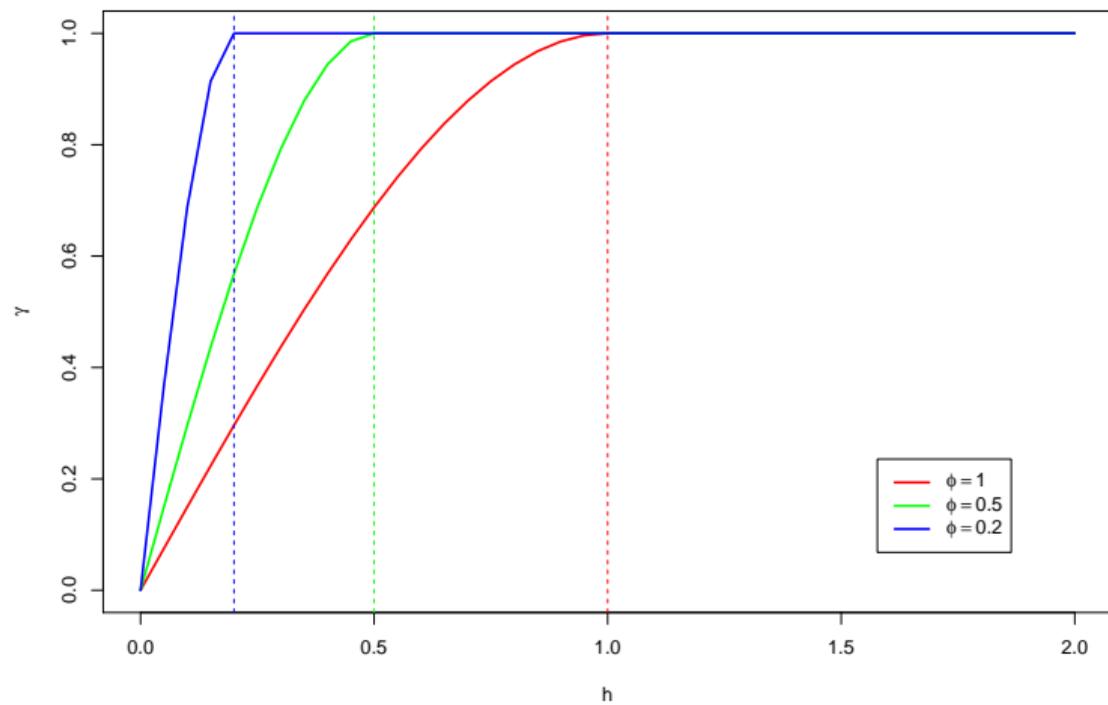
$$\rho(\mathbf{h}) = 1 - 1.5 \frac{\mathbf{h}}{\phi} + 0.5 \left(\frac{\mathbf{h}}{\phi} \right)^3 \text{ se } \mathbf{h} < \phi, 0 \text{ c.c..}$$

Em particular, o suporte compacto indica que para distâncias maiores que ϕ , o processo é independente.

Matérn

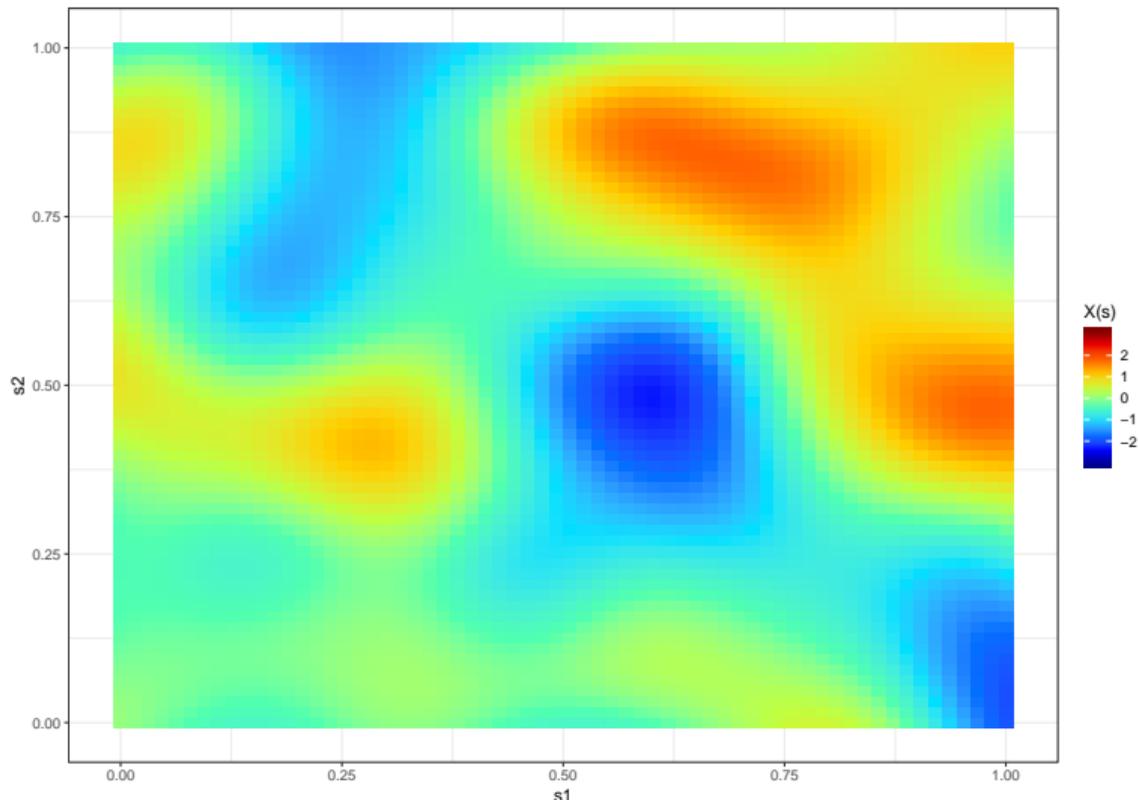


Esférico

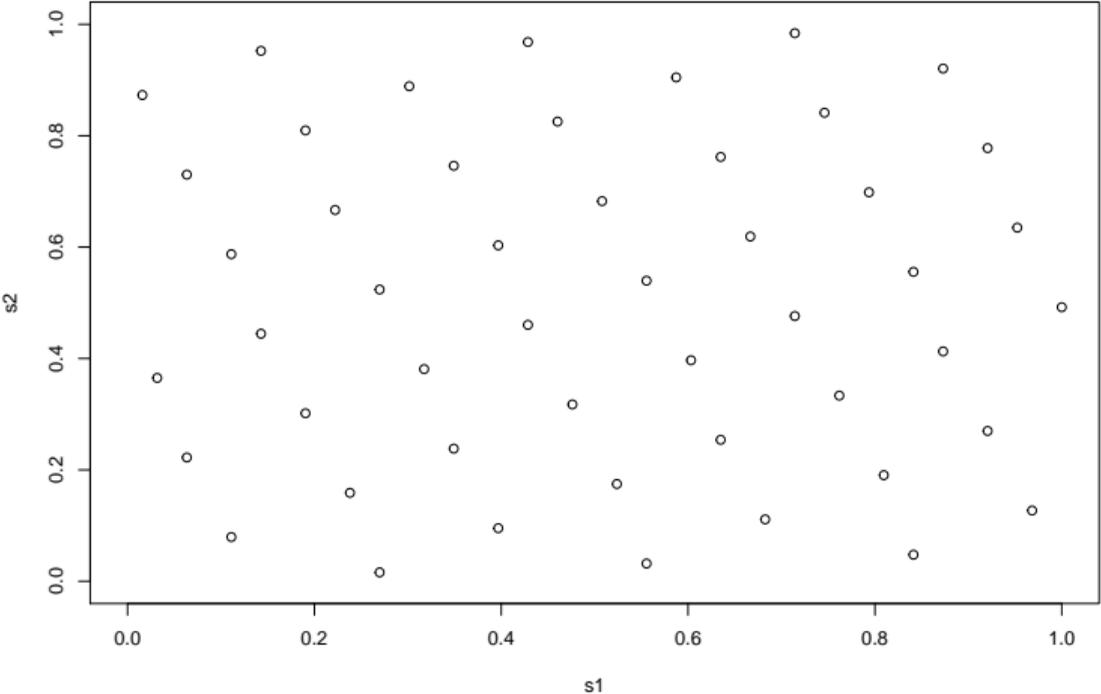


Amostras finitas

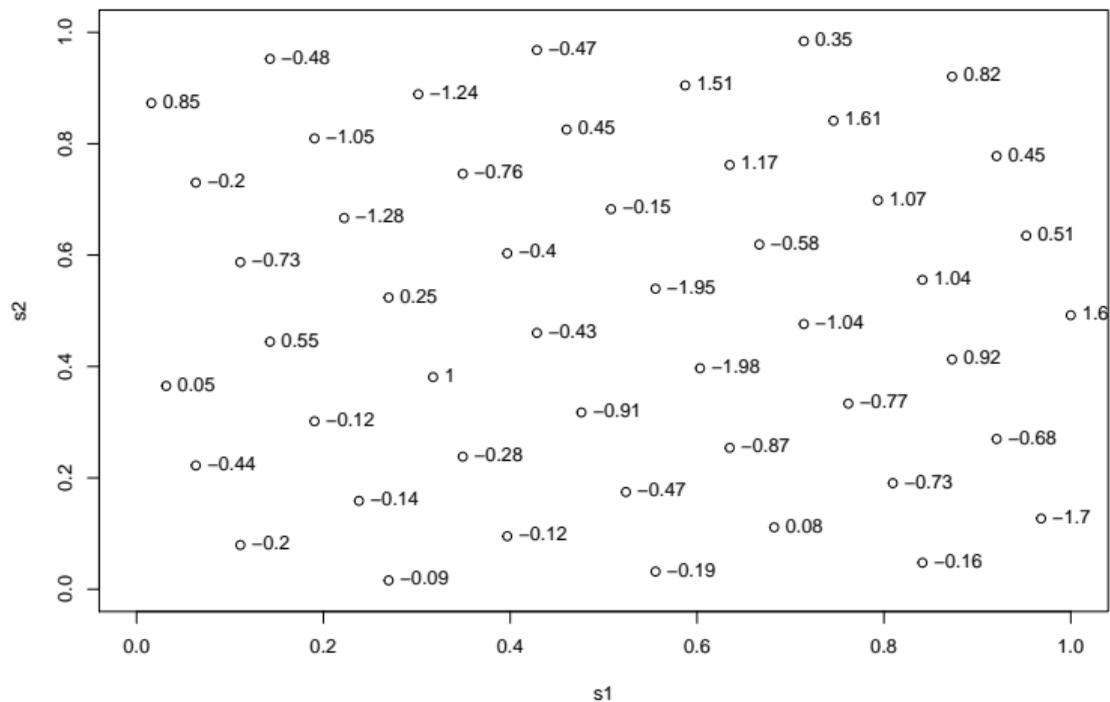
Como comentei na primeira aula: na prática, observamos o processo apenas em uma coleção finita de pontos. Por exemplo:



Amostras finitas



Amostras finitas



Variograma empírico

O variograma empírico é dado por

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j) \in N(h)} (x_i - x_j)^2,$$

onde

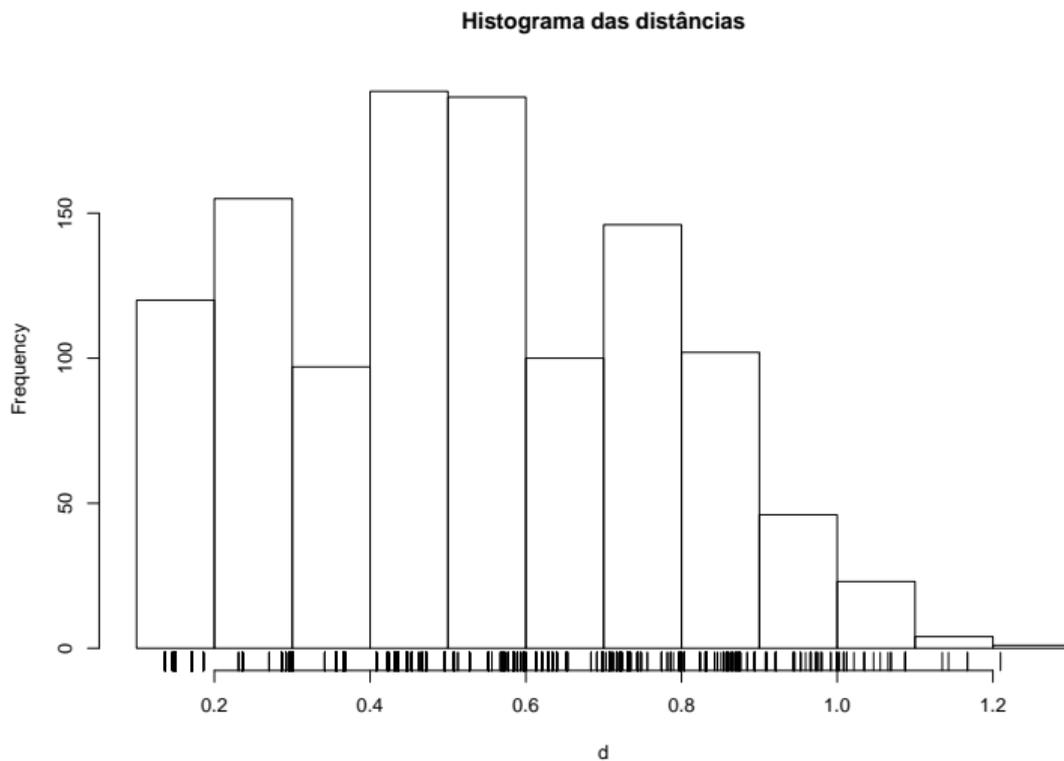
$$N(h) = \{(i, j) : h - \delta \leq \|\mathbf{s}_i - \mathbf{s}_j\| \leq h + \delta\},$$

e $h = \|\mathbf{h}\|$, $\delta < h$.

Você pode interpretá-lo como um histograma. δ , por exemplo, representa a largura dos anéis que envolvem \mathbf{s}_j a $h \pm \delta$ de distância da observação \mathbf{s}_i (e vice-versa).

Porém, você só terá informação (aproximada) sobre o variograma nas distâncias que amostrar o processo!

Distribuição de distâncias



Variograma empírico (+ variograma verdadeiro)

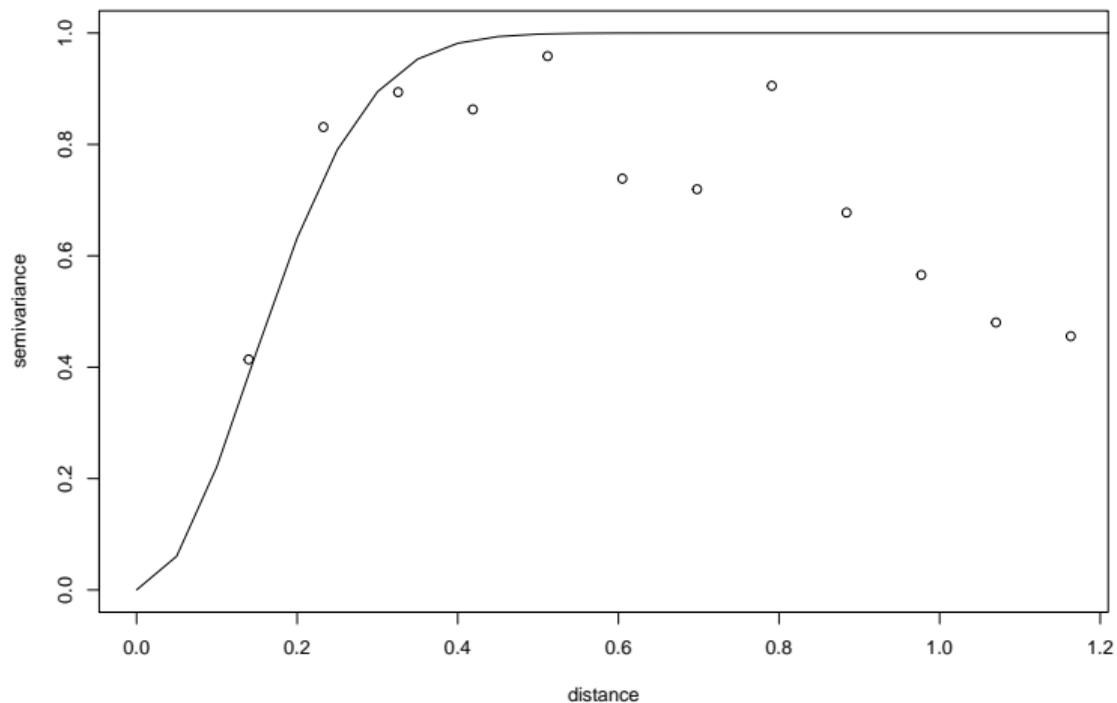
Código (no geoR)

```
v <- geoR::variog(coords = cbind(ss1, ss2), data = Z)
plot(v)
points(h, 1 - geoR::cov.spatial(h, cov.model = "gaussian",
                                cov.pars = c(sigma2 = 1, phi = 0.2)),
      type = "l")
```

Importante: não estamos estimando ainda! Considere o variograma “verdadeiro”...

Variograma empírico (+ variograma verdadeiro)

```
## variog: computing omnidirectional variogram
```



Resolução

O Cressie (1993) faz uma observação importante na página 112: seja o modelo

$$X(\mathbf{s}) = \mu(\mathbf{s}) + \eta(\mathbf{s}) + \varepsilon(\mathbf{s}).$$

- ▶ $\mu(\mathbf{s}) = \mathbb{E}(X(\mathbf{s}))$ é chamada de *large scale variation* ou variação determinística (veja nossa reflexão). Também são chamadas (em regressão) de efeitos ecológicos ou ambientais.
- ▶ $\eta(\mathbf{s})$ são GRFs com média zero, L^2 contínuos e intrínsecamente estacionários, cujo variograma é não-trivial para $\|\mathbf{h}\| \geq \min\{\|\mathbf{s}_i - \mathbf{s}_j\|\}_{1 \leq i < j \leq n}$. Esta é a chamada *small scale variation*, ou efeito aleatório espacial. Note que ele só é detectável a partir da *resolução* amostral.
- ▶ $\varepsilon(\mathbf{s})$ é a *micro-scale variation*, e inclui efeitos que não podem ser medidos na resolução amostral, além do erro instrumental.

Por isso: **sempre inclua um efeito de nugget no seu modelo.**

Referências I

Cressie, N. (1993). *Statistics for Spatial Data, 2nd edition*. Wiley, New York.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.