Métodos bootstrap para dados com dependência espacial.

Diogo Francisco Rossoni[12]; Vinícius Basseto Félix[1]

Resumo: Métodos bootstrap são largamente utilizados na estatística. Todavia, sua proposta original considera que os dados sejam independentes. Logo, a sua aplicação em dados que possuem algum tipo de dependência (temporal, espacial, espaço-temporal, etc) não é possível. Assim, este estudo tem o objetivo de apresentar 3 métodos bootstrap para a aplicação em dados com dependência espacial. Para verificar a eficácia destes métodos foram utilizados dados de profundidade (batimetria) de um trecho de 6 km do Rio Paraná coletados com auxílio de um sonar acoplado a um GPS no primeiro semestre de 2012 totalizando 450 pontos amostrais em transectos "lineares". O método de bootstrap da validação cruzada e o método de bootstrap da semivariância apresentaram resultados compatíveis com o método clássico de aproximação por máxima verossimilhança.

Palavras-chave: Bootstrap; dependência espacial; Geoestatística.

Abstract: Bootstrap methods are widely used in statistics. However, the original formulation of the bootstrap idea considers that data are independent. So, the bootstrap methods are not recommended for data with some kind of dependence. This paper has the aim to propose three bootstrap methods for data with spatial dependence. To evaluate this methods was used river depth data that was obtained from Paraná River, in a sector of 6 km. The cross-validation and variogram bootstrap methods presented similar results as one obtained for maximum likelihood approximation.

Key-words: Bootstrap; Spatial dependence; Geostatistics.

1 Introdução

Uma das características da Geoestatística é que na maioria dos casos a repetição de um experimento não é possível ou viável. Tome como exemplo um estudo que busque verificar a resistência a penetração no solo. Após ser feita a medição em

[1] Departamento de Estatística – Universidade Estadual de Maringá - UEM
[2] dfrossoni@uem.br

determinada posição, não é possível repetir esse experimento no mesmo local. Dessa forma, a grande maioria das aplicações da Geoestatística tem como finalidade obter estatísticas pontuais acerca da amostra em estudo.

O método de bootstrap foi proposto por Efron (1979) e, de uma forma simples, consiste em um procedimento de reamostragem com reposição. O método é largamente utilizada nas mais diversas áreas de estatística.

Em sua formulação original, o bootstrap é uma variação do método de Monte Carlo que numericamente determina a forma da distribuição (CHERNICK, 1999; DIACONIS, EFRON, 1983). Dada uma amostra aleatória e independente, o método consiste nas seguintes etapas (OLEA; PARDO-IGÚZQUIZA, 2010):

1. Obtenha aleatoriamente uma amostra com reposição da amostra original;
2. Calcule e salve as estatísticas de interesse;
3. Retorne a etapa 1 e repita o processo pelo menos 1000 vezes;
4. Apresente corretamente os valores obtidos para cada estatística;
5. Pare.

Todavia, na presença de correlação espacial a etapa 1 é inapropriada pois a reamostragem pode resultar em uma amostra contendo dados não correlacionados.

Uma das primeiras soluções para o problema de bootstrap com dados correlacionados foi proposta por Solow (1985). A ideia básica do trabalho foi transformar dados correlacionados em dados não correlacionados. Journel (1994) propôs um método de reamostragem condicional baseado na distribuição do erro. Berkowitz e Kilian (2007) apresentaram uma reamostragem bootstrap em bloco para dados com dependência temporal. Liang el al (2013) propuseram um método de reamostragem através de sub amostras, todavia, aplicável apenas em grandes bancos de dados. Outras abordagens podem ser verificadas em Olea e Pardo-Igúzquiza (2010) e Chrysikopoulos e Vogler (2004).

Cribari-Neto e Zarkos (2001, 2004) apresentaram uma série de vantagens de aplicação dos métodos bootstrap, desde a facilidade de obtenção de novas estimativas, construção de intervalos de confiança e correção de viés.

O objetivo deste trabalho foi propor três algoritmos para obtenção de novas amostras espacialmente correlacionadas através das técnicas de bootstrap. Estes algoritmos propostos são adaptações de técnicas bootstrap previamente utilizadas em regressões não-lineares e séries temporais. Além disso, tem-se o objetivo de construir intervalos de confiança para as estatísticas de interesse.

2 Material e métodos

Um processo espacial pode ser expresso pela soma de três componentes:

1. um componente estrutural, correspondente a um valor médio ou a uma tendência;
2. um componente aleatório, espacialmente correlacionado;
3. um ruído aleatório.

Assim, seja $Z(s_i)$ uma variável aleatória em que s_i denota uma posição em uma, duas ou mais dimensões $(s_i \in \mathbb{R}^n)$, então

$$Z(s_i) = \mu(s_i) + \varepsilon'(s_i) + \varepsilon''(s_i), \quad (1)$$

em que:

$\mu(s_i)$ é uma função determinística que representa o componente estrutural;

$\varepsilon'(s_i)$ é um termo estocástico que varia localmente e é espacialmente correlacionado;

$\varepsilon''(s_i)$ é um ruído aleatório, não correlacionado.

Os fenômenos espaciais geralmente apresentam as seguintes características:

- únicos, não replicáveis;
- definidos (geralmente) no domínio de 2 ou 3 dimensões;
- muito complexos para serem modelados por um processo determinístico;
- conhecidos por amostras tomadas em localizações distintas.

Várias medidas se prestam para a descrição dessa relação, tais como a autocovariância e autocorrelação, usuais na análise de séries temporais. Em Geoestatística, a medida normalmente utilizada é a semivariância. Ao contrário da covariância e correlação, a semivariância é uma medida de dissimilaridade, ou seja, o valor da semivariância é maior à medida que as variáveis estão menos associadas.

A semivariância é a medida do grau de dependência espacial entre duas amostras. A magnitude da semivariância entre dois pontos depende da distância entre eles. O gráfico das semivariâncias em função da distância *h* é chamado de

semivariograma. O semivariograma é uma ferramenta que permite representar quantitativamente a variação de um fenômeno regionalizado no espaço (JOURNEL, 1978).

O estimador clássico da semivariância, proposto por Matheron (1962) e apresentado por Cressie (1993) é dado pela equação,

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|}\sum_{N(h)}\left[Z(s_i) - Z(s_j)\right]^2, \quad (2)$$

em que:

$\hat{\gamma}(h)$ é o estimador da semivariância;

$N(h) \equiv \left\{(s_i, s_j) : s_i - s_j = h; i, j = 1,...,n\right\}$, isto é, s_i e s_j são posições espaciais separadas pela distância h, tal que $h \in \mathbb{R}^d$;

$|N(h)|$ é o número de pares de valores medidos, $Z(s_i)$ e $Z(s_j)$, separados pela distância *h;*

A sensibilidade dos semivariogramas para detectar a variabilidade espacial das amostras está diretamente ligada ao melhor ajuste de algum modelo teórico ao semivariograma empírico. Modelos teóricos de semivariogramas são superpostos à sequência de pontos obtidos no semivariograma empírico, de modo que a curva que melhor se ajusta aos pontos representa a magnitude, o alcance e a intensidade da variabilidade espacial da variável estudada. Os principais modelos abordados na literatura são: gaussiano, esférico e exponencial.

A partir do modelo teórico é possível, através do estimador de krigagem, obter o valor de um ponto não amostrado, isto é, considere um ponto s_0 não amostrado, tem-se que o estimador de krigagem visa obter $Z(s_0)$ através de:

$$\hat{Z}(s_0) = \sum_{i=1}^{n}\lambda_i Z(s_i) \quad (3)$$

em que:

n é o tamanho da amostra

$Z(s_i)$ são os valores observados e

λ_i são os pesos associados a cada valor observado.

A obtenção dos pesos λ_i pode ser feita de várias formas: krigagem ordinária, krigagem universal, etc.

Em todos os casos, o sistema de equações para a obtenção dos λ_i leva em consideração o modelo ajustado ao semivariograma empírico.

Os métodos bootstrap propostos neste trabalho terão enfoque em duas abordagens da Geoestatística: a semivariância e o preditor de krigagem.

2.1 Bootstrap da semivariância

O algoritmo para o bootstrap da semivariância é o mesmo definido por Davison e Hinkley (1997) para regressão não linear. Podemos escrever a semivariância como $\hat{\gamma}(h) = \gamma_{\text{mod}}(h) + \varepsilon$, em que $\gamma_{\text{mod}}(h)$ é um modelo de semivariância teórico estimado. O algoritmo para obter novas estimativas bootstrap de semivariância é definido como:

1- Considere $h^* = h$;

2- Amostre aleatoriamente e com reposição ε^* de $\varepsilon - \bar{\varepsilon}$;

3- A nova semivariância será $\gamma^*(h^*) = \gamma_{\text{mod}}(h^*) + \varepsilon^*$.

É possível utilizar a mesma abordagem dos modelos não lineares, pois o erro da semivariância é considerado iid $\sim N(0, \sigma^2)$.

Todavia, caso o interesse seja reamostrar a partir da amostra original – e não da semivariância – esse procedimento não é aplicável.

2.2 Bootstrap de validação cruzada

Podemos definir o erro de predição como $\varepsilon(s_i) = Z(s_i) - \hat{Z}(s_i)$, em que $\hat{Z}(s_i)$ pode ser obtido através do preditor de krigagem. Entretanto, esse preditor em particular apresenta um pequeno problema: a proposta inicial do preditor de krigagem é estimar um ponto não amostrado. Quando estimamos um ponto já amostrado através da krigagem, praticamente não erramos (erros na casa de 10^{-31}), logo, as novas amostras bootstrap acabam por não mostrar diferenças significativas da amostra original. A solução proposta nesse trabalho é criar um vetor de erros através da validação cruzada, isto é, apaga-se um ponto observado e estima-se ele a partir da krigagem considerando os demais pontos amostrados. Logo, o algoritmo fica definido como:

1- Considere $s_i^* = s_i$;

2- Obtenha $\hat{Z}(s_i)$ através de $\hat{Z}(s_i) = \sum_{j \neq i}^{n-1} \lambda_j Z(s_j)$;

3- Obtenha $\varepsilon = Z(s_i) - \hat{Z}(s_i)$;

4- Amostre aleatoriamente e com reposição ε^* de $\varepsilon - \bar{\varepsilon}$;

5- A nova amostra será $Z^*(s_i) = \hat{Z}(s_i) + \varepsilon^*$.

2.3 Reamostragem bootstrap em bloco.

A ideia básica da reamostragem bootstrap em blocos é dividir a malha amostral em blocos e rearranjar aleatoriamente – e sem reposição - esses blocos. A versão mais simples dessa ideia divide a amostra em b blocos não sobrepostos, cada um de tamanho l, em que $n = b.l$.

A nova malha é constituída pela realocação dos b blocos, cada qual com probabilidade b^{-1}. Vide exemplo na Figura 1.

Este método é uma adaptação da proposta inicial de reamostragem em bloco para séries temporais estacionarias (DAVISON; HINKLEY, 1997).

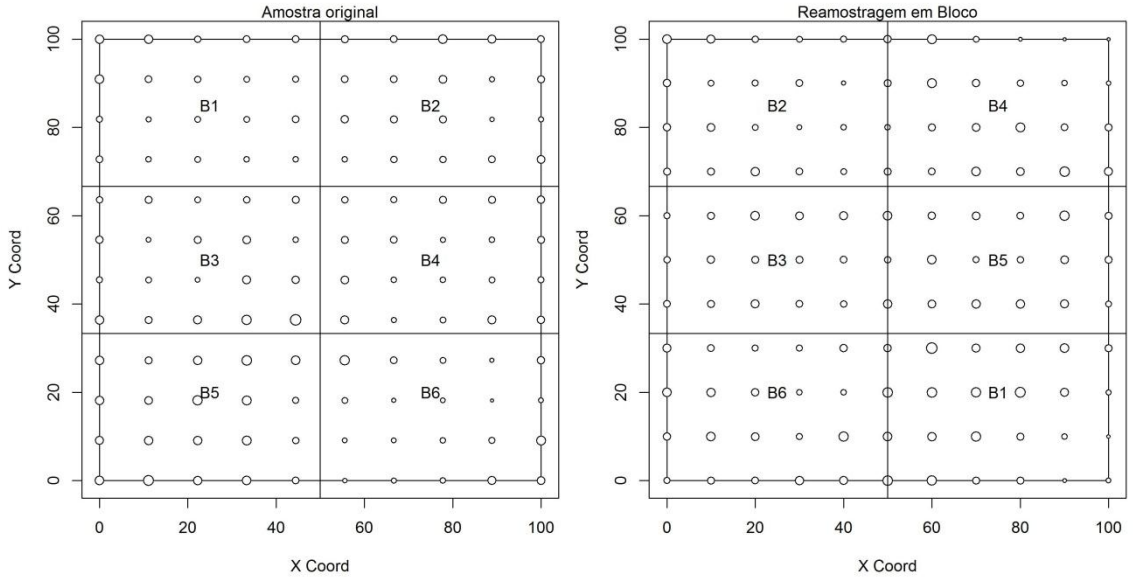

Figura 1 Exemplo de reamostragem em bloco considerando $b = 6$ e $l = 20$.

2.4 Dados experimentais.

Neste trabalho foram utilizados dados de profundidade de um trecho de 6 km do Rio Paraná coletados com auxílio de um sonar acoplado a um GPS no primeiro semestre de 2012 totalizando 450 pontos amostrais em transectos "lineares" (Figura 2).

Figura 2 **(a)** Área de estudo: trecho do Rio Paraná; **(b)** Pontos de profundidade georreferenciados amostrados utilizando um sonar acoplado a um GPS.

3 Resultados e discussões

A distância entre os pontos amostrais variou de $1{,}16m$ (mínima) a $6789{,}12m$ (máxima). A profundidade média foi de $4{,}41m$, sendo que a máxima profundidade amostral foi de $9{,}6m$ e a mínima de $1{,}5m$.

Procedeu-se com a estimação da semivariância através do estimador clássico de Matheron. O modelo ajustado foi:

$$\gamma_{\text{mod}}(h) = 4{,}1022.\left(1 - e^{-\left(\frac{h}{89{,}4070}\right)}\right) \qquad (4)$$

Que corresponde ao modelo esférico, com efeito pepita (C_0) 0, contribuição (C_1) 4,1022 e alcance (a) 89,4070 (Figura 3).

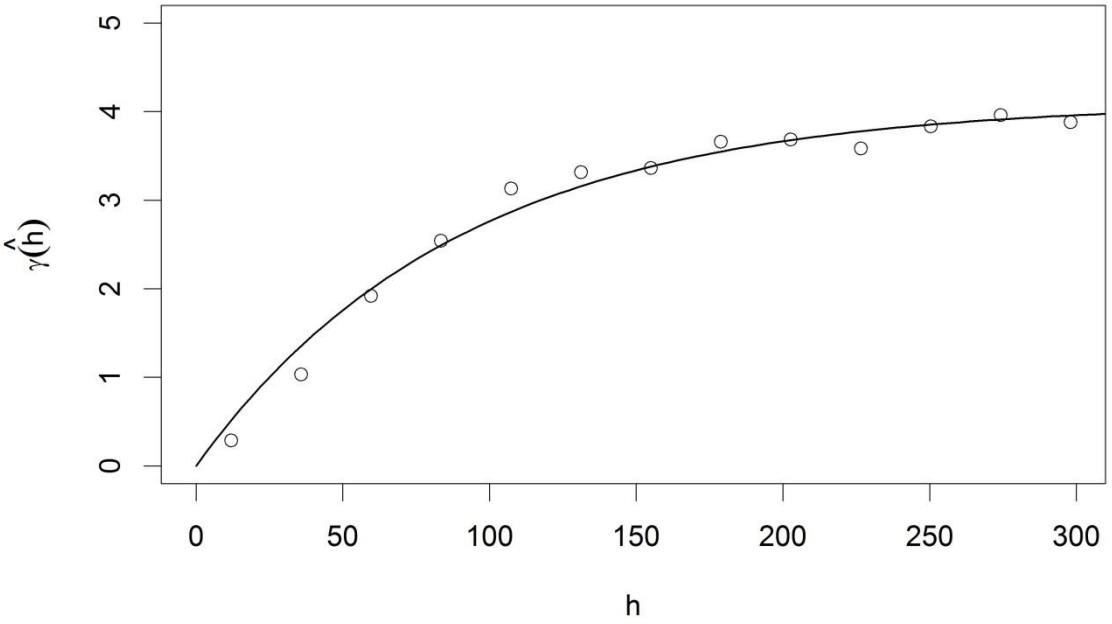

Figura 3 Modelo exponencial ajustado ao semivariograma empírico.

Para o cálculo das estimativas bootstrap dos parâmetros C_0, C_1 e a, cada um dos procedimentos bootstrap anteriormente descritos foi replicado 1000 vezes. Para a reamostragem bootstrap em bloco, foram considerados 5 blocos com 90 observações e 10 blocos com 45 observações. Todas as rotinas bootstrap implementadas estão no Anexo A.

A Figura 4 apresenta o boxplot para os parâmetros C_0, C_1 e a, obtidos pelo método bootstrap de validação cruzada e método bootstrap de regressão da semivariância. Percebe-se que, em geral, o método de validação cruzada apresenta maior amplitude das novas estimativas. Vale ressaltar que as estimativas originais estão contidas em todos os boxplots.

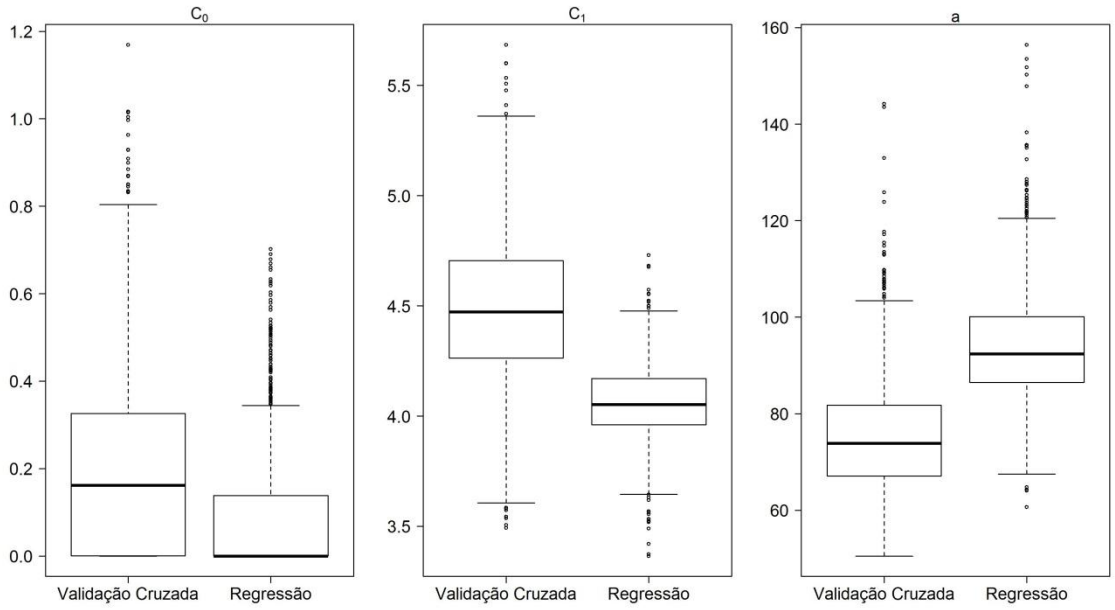

Figura 4 Boxplot das estimativas bootstrap pelo método de validação cruzada e método de regressão da semivariância.

A Figura 5 apresenta o boxplot para os parâmetros C_0, C_1 e a, obtidos pelo método de reamostragem bootstrap em bloco. Mesmo com a variação da quantidade de blocos, apenas para o parâmetro C_0 os resultados foram próximos aos obtidos pelos métodos de validação cruzada e de regressão da semivariância. Para os parâmetros C_1 e a houve uma superestimação dos parâmetros, sendo que é possível verificar valores para C_1 próximo de 8000 e valores para a próximo de 800000. Pressupõe-se que estes valores divergentes das estimativas originais ocorreram devido a um ajuste equivocado dos modelos teóricos as novas amostras bootstrap. Além disso, Davison e Hinkley (1997) comentam que a maior dificuldade para reamostragem em bloco é gerar novas amostras (malha) que são menos dependentes do que a amostra original. Isso pode acarretar em um conjunto de dados sem dependência espacial, o que resultaria em modelos de efeito pepita puro.

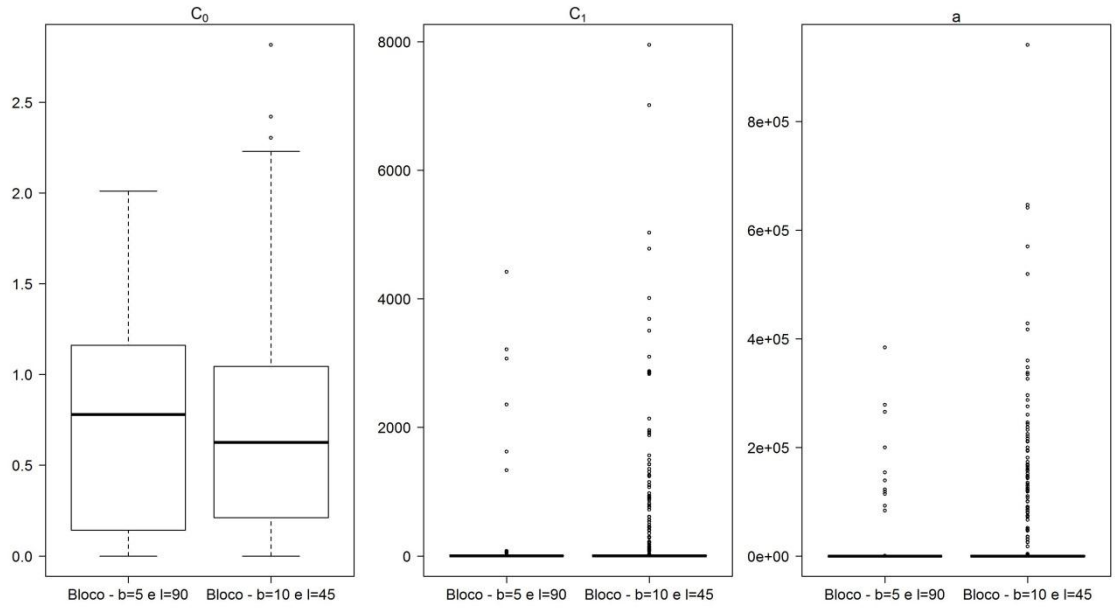

Figura 5 Boxplot das estimativas bootstrap pelo método da reamostragem em blocos

A partir das 1000 novas estimativas de C_0, C_1 e a, construiu-se intervalos de confiança Monte Carlo com 5% de significância (BUCKLAND, 1984).

A Tabela 1 apresenta as estimativas originais da amostra, bem como as obtidas pelos métodos bootstrap supracitados. Para critério de comparação, construiu-se os intervalos de confiança através da aproximação por máxima verossimilhança. Esses intervalos foram obtidos a partir do modelo ajustado (equação(4)) através do comando *confint* (R CORE TEAM., 2012).

Tabela 1 Intervalos de confiança com 5% de significância.

Parâmetros	Estimativas obtidas a partir da amostra original	Validação Cruzada	Regressão da semivariância	Bloco $(b=5)$ $(l=90)$	Bloco $(b=10)$ $(l=45)$	Aproximação por máxima verossimilhança (AMV)
C_0	0	(0; 0,76)	(0; 0,51)	(0; 1,97)	(0; 1,81)	(0; 0,19)
C_1	4,10	(3,83; 5,23)	(3,66; 4,39)	(0,28; 2356,20)	(0,04; 1256,78)	(3,83; 4,42)
a	89,40	(55,44; 105,56)	(74,63; 122,28)	(48,76; 199817,6)	(21,93; 211438)	(73,35; 109,47)

Pode-se verificar que os IC obtidos pelo método de Validação Cruzada, regressão da semivariância e AMV contiveram as estimativas originais. O método de Validação Cruzada apresentou a maior amplitude intervalar entre estes três.

Como já verificado na Figura 5, os IC obtidos pelo método de reamostragem em bloco não obtiveram bons resultados, apresentando os intervalos com maior amplitude e com superestimação de parâmetros (Tabela 1).

4 Conclusões

A aplicação de métodos bootstrap para dados com dependência espacial mostrou-se satisfatória. O método de bootstrap a partir da validação cruzada e da regressão da semivariância apresentaram intervalos de confiança compatíveis com os obtidos pela AMV. Os métodos bootstrap baseados em reamostragem em bloco obtiveram estimativas superestimadas.

Pretende-se desenvolver estudos futuros, através de simulação, para a comparação destes diversos métodos em vários cenários de dependência espacial.

5 Bibliografia

BERKOWITZ, J.; KILIAN, L. Recent developments in bootstrapping time series. **Econometric Reviews**, n. June 2013, p. 37–41, 2007. BERKOWITZ, J.; KILIAN, L. Recent developments in bootstrapping time series. **Econometric Reviews**, n. June 2013, p. 37–41, 2007.

BUCKLAND, S. T. Monte Carlo Confidence Intervals. **Biometrics**, v. 40, n. 3, p. 811–817, 1984.

CHERNICK, M. R. **Bootstrap methods : a practitioner's guide**. [s.l: s.n.].

CHRYSIKOPOULOS, C. V.; VOGLER, E. T. Estimation of time dependent virus inactivation rates by geostatistical and resampling techniques: application to virus transport in porous media. **Stochastic Environmental Research and Risk Assessment (SERRA)**, v. 18, n. 2, p. 67–78, 1 abr. 2004.

CRESSIE, N. **Statistics for spatial data**. New York: [s.n.].

CRIBARI-NETO, F.; ZARKOS, S. G. Heteroskedasticity-consistent covariance matrix estimation:white's estimator and the bootstrap ∗. **Journal of Statistical Computation and Simulation**, v. 68, n. 4, p. 391–411, mar. 2001.

CRIBARI-NETO, F.; ZARKOS, S. G. Leverage-adjusted heteroskedastic bootstrap methods. **Journal of Statistical Computation and Simulation**, v. 74, n. 3, p. 215–232, mar. 2004.

DIACONIS P, E. B. Computer-intensive methods in statistics. **Sci Am**, v. 5, p. 116–130, 1983.

EFRON, B. Bootstrap methods: another look at the jackknife. **The Annals of Statistics**, v. 1, p. 1–26, 1979.

JOURNEL, A. G. Resampling from stochastic simulations. **Environmental and Ecological Statistics**, v. 1, n. 1, p. 63–91, mar. 1994.

JOURNEL, A. G. **Mining geostatistics**. London: Academic Press, 1978.

LIANG, F. et al. A Resampling-Based Stochastic Approximation Method for Analysis of Large Geostatistical Data. **Journal of the American Statistical Association**, v. 108, n. 501, p. 325–339, mar. 2013.

MATHERON, G. **Traité de géostatistique appliquée. Mémoires du Bureau de Recherches Géologiques et Minieres**. [s.l.] Editions Technip, 1962. v. 14

OLEA, R. A.; PARDO-IGÚZQUIZA, E. Generalized Bootstrap Method for Assessment of Uncertainty in Semivariogram Inference. **Mathematical Geosciences**, v. 43, n. 2, p. 203–228, 24 fev. 2010.

R CORE TEAM. **R: A language and environment for statistical computing.**VienaR Foundation for Statistical Computing, , 2012. Disponível em: <http://www.r-project.org/.>

SOLOW, A. R. Bootstrapping correlated data. **Journal of the International Association for Mathematical Geology**, v. 17, n. 7, p. 769–775, out. 1985.

6 Anexo A

```
######################################################################
#CARREGAMENTO DO PACOTE
######################################################################
rm(list=ls())
library(geoR)
######################################################################
#IMPORTAÇÃO
######################################################################
data<-read.geodata("dados.txt",h=T)
points(data)
summary(data)
```

```
#####################################################################
#TIPOS DE BOOTSTRAP
#####################################################################
B<-1000 ##Bootstrap
parametros<-data.frame(C0=rep(0,B),C1=rep(0,B),a=rep(0,B))
parametros.block<-data.frame(C0=rep(0,B),C1=rep(0,B),a=rep(0,B))
parametros.reg<-data.frame(C0=rep(0,B),C1=rep(0,B),a=rep(0,B))
#####################################################################
#PLOT  DO SEMIVARIOGRAMA
#####################################################################
var.or<-variog(data,max.dist=310)
plot(var.or,ylim=c(0,5),xlab=expression(h),ylab=expression(hat(gamma(h))))
mod.or<-variofit(var,ini.cov.pars=c(3.5,142),nugget=0.375, cov.model="exp")
lines(mod.or,lwd=2)
#####################################################################
#VALIDAÇÃO CRUZADA
#####################################################################
v.cruzada<- xvalid(data, model=mod.or)
ko<-krige.conv(data,location = data$coords, krige=krige.control(obj.model = mod.or))
erro<-v.cruzada$error

#####################################################################
#BOOTSTRAP - VALIDAÇÃO CRUZADA
#####################################################################
for(i in 1:B){
  erro.new<-sample((erro-mean(erro)),length(erro),replace = T)
  new.data<-ko$predict+erro.new
  sample<-as.geodata(data.frame(data$coords[,1],data$coords[,2],new.data))
    var<-variog(sample,max.dist=300)
  #plot(var,main="Clássico",ylim=c(0,5),xlab="Distância",ylab="Semivariância")
  mod<-variofit(var,ini.cov.pars=c(3.5,142),nugget=0.375, cov.model="exp")
   #lines(mod)
  parametros[i,]<-as.numeric(c(mod$nugget,mod$cov.pars))}
#####################################################################
```

```
#BOOTSTRAP - BLOCO
#####################################################################
for(i in 1:B){
 l<-5 #Size of the block
 block<-matrix(data$data,ncol = l)
 id<-sample(1:l,size = l,replace = F)
 block.data<-as.numeric(block[,id])
 sample<-data
 sample$data<-block.data
 var<-variog(sample,max.dist=300)
 #plot(var,main="Clássico",ylim=c(0,5),xlab="Distância",ylab="Semivariância")
 mod<-variofit(var,ini.cov.pars=c(3.5,142),nugget=0.375, cov.model="exp")
 #lines(mod)
 parametros.block[i,]<-as.numeric(c(mod$nugget,mod$cov.pars))}
#####################################################################
# REGRESSÃO DA SEMIVARIÂNCIA
#####################################################################
x<-var.or$u
y<-var$v
#plot(y~x)
mod<-nls(y~c0+c1*(1-exp(-x/a)),start=list(c0= mod.or$nugget,c1= mod.or$cov.pars[1]
,a= mod.or$cov.pars[2]), algorithm = "port", lower=c(c0=0,c1=0,a=0 ))
#lines(x = x,y=predict(mod))
erro<-y-predict(mod)


#####################################################################
#BOOTSTRAP - REGRESSÃO DO SEMIVARIÂNCIA
#####################################################################
for(i in 1:B){
 erro.new<-sample((erro-mean(erro)),length(erro),replace = T)
 y.new<-predict(mod)+erro.new
 plot(y.new~x)
 mod.new<-nls(y.new~c0+c1*(1-exp(-x/a)), start=list(c0= mod.or$nugget,c1=
```

```
mod.or$cov.pars[1],a= mod.or$cov.pars[2]), algorithm = "port", lower=c(c0=0,c1=0,a=0 ))
 s<-summary(mod.new)
 parametros.reg[i,]<-as.numeric(c(s$coefficients[1,1],s$coefficients[2,1],s$coefficients[3,1]))
}
#####################################################################
#INTERVALO DE CONFIANÇA DE MONTE CARLO
#####################################################################
par<-sort(parâmetros$c0) ###distribuição do parâmetro de interesse
m<-length(par)
CC <- .95 #Defines the confidence coefficient (95% here).
tail <- (1-CC)/2 #Computes the tail probability (.025 here).
lownum1 <- trunc(tail*m+.5) #Computes the obs. below the 2.5th percentile.
lownum2 <- ceiling(tail*m+.5) #Computes the obs. above the 2.5th percentile.
lp <- tail*m+.5-lownum1 #Computes the weight for lownum1.
upnum1 <- trunc((1-tail)*m+.5) #Computes the obs. below the 97.5th percentile.
upnum2 <- ceiling((1-tail)*m+.5) #Computes the obs. above the 97.5th percentile.
up <- 1-lp #Computes the weight for upnum1.

lower <- (1-lp)*par[lownum1]+lp*par[lownum2] # Lower 95% CI limit for mean.
upper <- (1-up)*par[upnum1]+up*par[upnum2] # Upper 95% CI limit for mean.
cat("Monte Carlo CI for Mean = (", # Print the limits of the CI,
   round(lower,4),round(upper,4),") \n") #rounded to 4 decimal places.
to 4 decimal places.
```