UNIVERSIDADE DE SÃO PAULO
ESCOLA SUPERIOR DE AGRICULTURA "LUIZ DE QUEIROZ"

PLANO DE DISSERTAÇÃO

CURSO DE PÓS-GRADUAÇÃO EM: Estatística e Experimentação Agronômica

NÍVEL: Mestrado

CANDIDATO: Bruno Henrique Fernandes Fonseca

ORIENTADOR: Prof. Dr. Paulo Justiniano Ribeiro Jr.

TÍTULO: Modelos espaço-temporais com resposta multivariada

PIRACICABA
2007

SUMÁRIO

1 INTRODUÇÃO

A modelagem estatística é um conjunto de ferramentas muito importante em diversos campos do conhecimento, que utilizam essas técnicas para tentar descrever o comportamento de um ou mais atributos que não possuem um modelo determinístico. De uma forma geral, esse tipo de estudo tenta explicar, o máximo possível, a variabilidade dos processos estocásticos através de uma ou mais variáveis explanatórias e a parte da variabilidade que não é captada pelo modelo é atribuída ao acaso proveniente do processo amostral envolvido, essa variação ao acaso recebe o nome na literatura de erro aleatório ou ruído branco.

Os primeiros modelos estatísticos propostos foram os lineares univariados, que assumem erros aleatórios independentes e identicamente distribuídos de uma distribuição de probabilidade gaussiana, além disso, todas as variáveis explanatórias eram consideradas fixas, ou seja, não existem distribuições de probabilidades associadas às covariáveis. No entanto, essas simplificações não são válidas na maioria dos processos naturais, logo, surgiu a necessidade de desenvolver técnicas mais sofisticadas para tentar modelar processos que possuem estruturas mais complexas de variabilidade.

Um campo de pesquisas que teve grande evolução nos últimos tempos foi a estatística espacial, que é formada por três grandes áreas de estudo: geoestatística, dados de área e processos pontuais, que são utilizadas conforme o tipo de dado em questão, este plano irá considerar apenas a primeira. A geoestatística é um conjunto de técnicas que tenta modelar um ou mais atributos que possuem localizações espaciais e pontuais conhecidas, sendo assim, essas ferramentas são úteis para capturar a correlação entre as observações dos atributos sob estudo, onde existe uma forte suspeita de que pontos espaciais mais próximos possuem valores observados dos atributos mais parecidos, ou seja, a estrutura de correlação entre as observações do processo estocástico é determinada através das distâncias entre os pontos espaciais amostrados.

Os estudos de séries temporais e dados longitudinais são outros campos de pesquisas que evoluíram muito com a sofisticação das estruturas dos modelos estatísticos. Onde, há algum tempo, é possível modelar as estruturas de correlação com um ou mais atributos ao longo do tempo, ou seja, assim como em geoestatística, se espera que observações mais próximas, agora no tempo, possuem valores dos atributos mais parecidos.

Diversos estudos atuais tentam modelar conjuntamente estruturas de correlações espaciais e temporais, na literatura esse tipo de ferramenta estatística é denominada de modelagem espaço-temporal.

Como dito acima, as pesquisas de diversas áreas podem possuir mais de uma variável resposta de interesse, se esses atributos sob estudo forem independentes deve-se propor um modelo estatístico para cada um deles, no entanto, se há evidências de que esses processos sejam dependentes, modelos multivariados devem ser propostos, ou seja, os modelos estatísticos devem capturar ao máximo a correlação entre as variáveis respostas, para tal, algumas técnicas têm sido utilizadas, assim como, distribuições de probabilidades conjuntas e cópulas.

Com a evolução computacional das últimas décadas, mistura de diversas técnicas têm sido implementadas, neste contexto, um exemplo para aplicação de técnicas de modelos espaço-temporais multivariados são os estudos de agricultura de precisão, onde se tem, por exemplo, interesse em estudar as propriedades de solo em uma determinada região, que é medida através de duas variáveis resposta de interesse dos pesquisadores ao longo de um determinado tempo, sendo que essas variáveis respostas são fortemente correlacionadas, sendo assim, após estabelecer toda a estrutura de correlação entre as variáveis é possível começar a coletar cada vez menos informações sobre uma das variáveis que seja mais cara, barateando, assim, os custos desses tipos de pesquisa, que em alguns casos se arrastam por muitos anos.

2 OBJETIVO

a) Fazer revisão bibliográfica sobre campos aleatórios;

b) Fazer revisão bibliográfica sobre campos aleatórios espaço-temporais;

c) Fazer revisão bibliográfica sobre modelos geoestatísticos multivariados;

d) Fazer revisão bibliográfica sobre cópulas;

e) Propor formas de modelar as estruturas de correlações espaciais, temporais e entre as variáveis.

3 REVISÃO BIBLIOGRÁFICA

Neste tópico serão apresentadas algumas técnicas e conceitos de modelagem espaço-temporal, de modelos multivariados e de técnicas de estimação.

3.1 Campos Aleatórios

Um campo aleatório ou função aleatória é algum atributo sob estudo que existe em algum espaço real d-dimensional, geralmente bi ou tri-dimensional, os valores reais da função aleatória não são conhecidos, sendo assim, é necessário fazer uma amostragem de pontos espaciais e nas localizações amostradas o atributo de interesse será medido, com os valores observados pode-se propor algum modelo ao problema e fazer predições aos pontos não observados, abaixo segue a notação de campos aleatórios:

$$\{Z(s) : s \in G \subset R^d\}, \tag{1}$$

sendo $Z(s)$ a notação para o campo aleatório na localização s do espaço sob estudo G.

Segundo Schmidt e Sansó (2006) e Le e Zidek (2006), a descrição de um campo aleatório é obtida através das distribuições acumuladas finito-dimensionais F se, para qualquer conjunto de pontos nas localizações s_1,s_2,...,s_n pertencentes à região G e qualquer inteiro n:

$$F_{S_1,S_2,...,S_n}(z_1, z_2, ..., z_n) \equiv P(Z(s_1) \leq z_1, Z(s_2) \leq z_2, ..., Z(s_n) \leq z_n)$$

Uma das distribuições de probabilidades mais utilizadas na literatura é a gaussiana, que devido as suas propriedades, é relativamente, a mais fácil de estabelecer distribuições conjuntas e fazer inferências. Sendo assim, um processo espacial, como em (1) é dito ser gaussiano se $Z(s)$ segue uma distribuição Normal n-variada, logo, devido às propriedades da distribuição em questão, a cada atributo $Z(s_i)$ é associada uma distribuição normal univariada.

Como um processo gaussiano segue uma distribuição normal n-variada, ele é completamente especificado pelo vetor de médias e pela matriz de variâncias e covariâncias. O vetor de médias é especificado pela presença ou ausência de covariáveis ao processo, se for considerado que não existe nenhuma tendência sobre a média do processo é dito que o

processo gaussiano possui média constante, e o vetor de médias possui n valores iguais, por outro lado, se há evidências de que existe alguma tendência na média do processo devida à presença de covariáveis, é necessário propor algum modelo estatístico para capturar essa tendência, o que aumenta o número de parâmetros a estimar. A matriz de variâncias e covariâncias deve ser positiva definida, o que não é de fácil elaboração, por conta disso, algumas funções de correlação já conhecidas, que produzem matrizes positiva definidas, são muito utilizadas na prática.

Em uma pesquisa de geoestatística não é possível ter mais de uma realização do processo devido aos custos envolvidos ou outros problemas,sendo assim, outras suposições devem ser impostas para que seja possível a realização de inferências. Na literatura e na prática a restrição mais utilizada é que o processo é estacionário, ou seja, a distribuição da função aleatória não depende da grandeza de escala das coordenadas, sendo assim, a distribuição conjunta dos $(Z(s_1), Z(s_2), ..., Z(s_n))$ é igual a distribuição conjunta de $(Z(s_1 + h), Z(s_2 + h), ..., Z(s_n + h))$, para qualquer incremento h.

Outra definição menos restritiva é que uma função aleatória $Z(s)$ é dita ter estacionariedade fraca se $E[Z(s)] = \mu$ e $Cov[Z(s), Z(s + h)] = C(h)$.

Com as definições acima, se tem que a média é igual em toda a região sob estudo e que a covariância entre atributos em locais diferentes só depende da distância entre os mesmos. Esse tipo de estacionariedade é conhecido na literatura como estacionariedade fraca ou de segunda ordem, uma observação importante é que a primeira restrição implica na segunda, no entanto, o contrário não é válido, a não ser que o processo espacial seja gaussiano, que produz equivalência entre as duas restrições. No entanto, nem sempre é fácil verificar as restrições de estacionariedade forte ou fraca, logo, outra possibilidade menos restritiva é assumir que os incrementos $[Z(s) - Z(s + h)]$ possuem estacionariedade. Esta característica é denominada de estacionariedade intrínseca (SCHANBENBERGER; GOTWAY, 2005). Sendo assim, um campo aleatório é dito ser intrinsecamente estacionário se $E[Z(s)]=\mu$ e $Var[Z(s) - Z(s + h)] = 2\gamma(h)$, em que $\gamma(h)$ é denominado semivariograma, e a relação $\gamma = C(0) - C(h)$ é válida.

Por conta da relação entre as covariâncias e o semivariograma, a variabilidade de campos aleatórios intrinsecamente estacionários pode ser estudada por qualquer uma das

medidas da relação, comumente na literatura de geoestatística utiliza-se o semivariograma.

Outra abordagem que pode ser adotada é quando o processo não possua nenhum tipo de estacionariedade, ou seja, ou a média varia ao longo da região sob estudo ou a variância não é constante. Como dito anteriormente, quando a média não apresenta constância em toda a região sob estudo, algum modelo pode ser proposto para capturar essa variação, geralmente adota-se modelos lineares com as coordenadas como covariáveis. Com relação a variância e covariâncias não constantes no campo aleatório, pode-se tentar uma abordagem mais simples com uma transformação nos dados originais, geralmente a família de transformações de Box-Cox é utilizada. No entanto, devido a complexidade do problema, pode-se adotar outras abordagens como modelos de deformações espaciais (SAMPSON; GUTTORP, 1992, SCHMIDT; O'HAGAN, 2003) ou convoluções espaciais (HIGDON, 2002, FUENTES; SMITH, 2001). Uma outra característica que pode surgir em eventos da natureza é que o processo possui algum tipo de estacionariedade, mas mesmo assim a função de covariância depende da direção, assim a função aleatória será considerada anisotrópica, ou seja, a variabilidade é constante em todo o campo, porém há diferenças nas correlações conforme a direção em que a distância está, esse tipo de problema é muito comum em estudos de poluentes na atmosfera, onde a direção dos ventos gera uma distorção na correlação entre pontos com a mesma distância. Quando existe anisotropia nos dados, basta incluir mais parâmetros na estrutura de correlação, não há dificuldade em modelar esse problema, porém a identificação de tal padrão a partir dos dados não é fácil. A forma mais comum de tratar a anisotropia é fazer transformações nos sistemas de coordenadas, utilizando geometria para tal. Na literatura geoestatística quando um processo possui anisotropia ele é chamado de processo estacionário heterogêneo e caso contrário processo estacionário homogêneo.

3.1.1 Propriedades da função de covariância

Encontrar funções que estabeleçam a estrutura de covariância, que seja válida, não é trivial, ou seja, não é fácil achar funções de correlação que possuam um comportamento empírico, onde se espera que quanto maior a distância entre os pontos menor a correlação entre os atributos e que sejam positivas definida, logo, na literatura existem algumas famílias de funções de covariâncias que são conhecidamente válidas. No entanto antes de apresentar

tais famílias de funções válidas, seguem as propriedades das funções de covariância de um processo estacionário de segunda ordem:

(i) $Cov[Z(s), Z(s+0)] = Var[Z(s)] = C(0) \geq 0$;

(ii) $C(h) = C(-h)$;

(iii) $C(0) \geq \mid C(h) \mid$;

(iv) $C(h) = Cov[Z(s), Z(s+h)] = Cov[Z(0), Z(h)]$;

(v) Se C_j são funções de covariância válidas, então $\sum_j b_j \ C_j(h)$ e $\prod_j \ C_j(h)$ são funções válidas;

(vi) Se $C(h)$ é válida para um espaço d-dimensional, então $C(h)$ é válida para todo espaço menor que d.

3.1.2 Famílias de funções de covariância

A Família Matérn

Essa família de correlações foi proposta por Berfil Matérn (1986) e possui a seguinte função:

$$C(h) = 2^{\kappa} - \Gamma(\kappa)^{-1}(h/\phi)^{\kappa} K_{\kappa}(h/\phi),$$

os parâmetros dessa função são $\phi > 0$ e $\kappa > 0$, que são vinculados a escala com a dimensão de distância e suavidade do processo e $K_{\kappa}(.)$ é a função Bessel de ordem κ.

A Família Exponencial Potência

$$C(h) = \exp{(h/\phi)^{\kappa}},$$

essa família também possui dois parâmetros, com mesmas interpretações da família Matérn, no entanto agora κ limitado no intervalo $[0, 2]$. Cabe ressaltar que a família Matérn com κ=1/2 é igual a função exponencial com κ=1.

Na literatura de geoestatística existem diversas outras famílias que são válidas.

3.1.3 Estimação dos Parâmetros

Estabelecidas às estruturas paramétricas, o próximo passo é fazer a estimação dos parâmetros. Se o campo aleatório é intrinsecamente estacionário pode-se trabalhar uma

estimação para o semivariograma, abaixo segue a expressão de uma estimativa empírica para o semivariograma através dos estimadores de momentos:

$$\widehat{\gamma}(h) = (\sum_{|N(h)|} (Z(s_i) - Z(s_j))^2)/2|N(h)| \tag{2}$$

em que $|N(h)|$ é o número de pontos abrangidos pela distância h. Se o processo for intrinsecamente estacionário o estimador de (2) é não viesado para γ. Já se o campo aleatório possui estacionariedade fraca o estimador do semivariograma informa sobre a função de correlação do processo, ou seja, a relação $\gamma = C(0) - C(h)$ estabelece a relação entre a função de correlação e o semivariograma.

Devido a relação entre o semivariograma empírico e as funções de correlação válidas, muitos trabalhos aplicados de geoestatística utilizam um modelo em função dos parâmetros da função de correlação que se ajuste aos valores do semivariogramas empíricos, isso pode ser feito através do método de mínimos quadrados ponderados ou por meio de métodos "AD-HOC", no entanto, essas duas abordagens para estimar os parâmetros da função de correlação podem não ser muito precisos, pois os valores dos semivariogramas empíricos podem se afastar muito do semivariograma real e desconhecido , devido ao tamanho e acaso amostral.

Por outro lado, se assume-se que o campo aleatório possui estacionariedade forte, pode-se optar por estimadores de Máxima verossimilhança e Máxima Verossimilhança Restrita, que se baseiam no logaritmo da função de verossimilhança abaixo:

$$l(\theta; Z_1, Z_2, ..., Z_n) = -(\ln(|\Sigma(\theta)|) + n\ln(2\pi) + (Z(s) - 1\mu)^t \Sigma(\theta)^{-1}(Z(s) - 1\mu))/2. \tag{3}$$

A função de verossimilhança em (3) é baseada na distribuição gaussiana associada ao campo aleatório. Os estimadores de máxima verossimilhança são muito utilizados por conta das suas propriedades assintóticas. No entanto em geoestatística, como tem-se distribuição normal multivariada associada aos campos aleatórios, a matriz de variância e covariância pode possuir dimensão muito grande, o que gera elevado tempo computacional ou até inviabiliza a estimação dos parâmetros. Uma abordagem muito utilizada nesse contexto é a Bayesiana, o que requer métodos computacionais intensivos, como por exemplo MCMC.

3.1.4 Krigagem

A krigagem nada mais é do que o processo de predição para os valores do campo aleatório que não foram amostrados. Ou seja, na prática os pesquisadores amostram alguns pontos dentro do espaço de interesse e realizam medição do atributo sob estudo. No entanto, existe interesse em conhecer como o atributo se comporta em todo o espaço, logo, utilizando os valores estimados para os parâmetros do modelo estabelecido é possível descrever o campo aleatório predito, contínuo no espaço. O nome krigagem é uma homenagem ao pesquisador sul-africano D.G. Krige que foi um dos pioneiros em estudos de predição espacial.

Os dois tipos de krigagem mais encontrados na literatura são a ordinária e simples, que se diferenciam quanto a pressuposição ou não de conhecimento sobre os parâmetros. Os dois tipos de krigagem são baseados nos estimadores de mínimos quadrados. Todos os valores observados são função da média do campo aleatório com uma correção gerada pelos parâmetros de variância e covariância do campo aleatório. Segue a expressão para a krigagem simples:

$$\rho(z;s) = \mu_z + \sigma^2 r' \Sigma^{-1}(z - \mu_z),$$

sendo $\rho(z;s)$ o valor predito para o campo aleatório sob estudo na posição s do espaço de interesse, μ é a média de Z, σ é a variabilidade devida ao processo espacial, r' é o valor da correlação e Σ a matriz de variância e covariância do campo aleatório Z.

Diggle e Ribeiro (2006) e Elmatzoglou (2006), discutem com mais detalhes os métodos e propriedades dos processos de krigagem.

3.2 Campos aleatórios espaço-temporais

Um atributo medido em diversas localizações dentro de uma região finita de estudo e ao longo de algum intervalo de tempo é fonte de discussão para três abordagens diferentes de modelagem do atributo de interesse. A primeira análise pode ser feita somente com características espaciais para cada tempo amostrado, ou seja, existirá um modelo geoestatístico para cada tempo observado, a segunda abordagem é estudar apenas o comportamento temporal de algum atributo em cada localização, assim para cada coordenada do espaço sob estudo será proposto um modelo utilizando técnicas de séries temporais ou dados longitudinais. A última abordagem, que é foco dessa subseção, estuda os dois comportamentos em

conjunto, ou seja, agora o modelo proposto deve capturar a correlação espacial, temporal e possivelmente cruzada, nessa abordagem conjunta, freqüentemente as análises marginais de tempo e espaço são utilizadas como exploração do conjunto de dados, o que pode fomentar muitas análises e interpretações do modelo final proposto.

Segue a notação de campo aleatório espaço-temporal:

$$\{Z(s,t) : s \in R^d, t \in R\}, \tag{4}$$

agora o campo aleatório Z depende do s, que é análogo ao de campo aleatório, e do t que é unidimensional e pertence ao R, logo o domínio do processo é R^d X R.

Porém, tempo e espaço de campos aleatórios definidos como em (4) podem não ser diretamente comparáveis, pois possuem escalas diferentes e os modelos propostos devem considerar essa característica. A distância euclidiana é diretamente aplicável a esse tipo de abordagem, uma vez que não leva em consideração a escala das diferenças de espaço e de tempo.

A modelagem espaço-temporal pode ser tratada de duas formas, utilizando técnicas de geoestatística ou de modelos estocásticos, a primeira abordagem é mais simples se for considerado que o campo aleatório possui distribuição gaussiana multivariada, o que exige estacionariedade do campo aleatório.

Agora, deve-se pensar em quatro especificações do campo aleatório, quanto a sua distribuição de probabilidade, quanto a sua estacionariedade espaço-temporal, quanto a separabilidade do espaço e do tempo e quanto a simetria da função de covariância.

Uma função aleatória é dita ser estacionária no espaço e no tempo, se a média do processo não depender de nenhuma covariável, sejam elas as coordenadas do tempo ou espaço ou outro fator associado ao estudo, além disso a matriz de variância e covariância só depende de separações do domínio do processo. Sendo assim, a $E(Z(s,t))=\mu$, para todo s e t, e a $Cov(Z(s_i,t_i),Z(s_j,t_j))$ só depende das distância $h = s_i - s_j$ e $u = t_i - t_j$.

Se o campo aleatório possuir estacionariedade espaço-temporal, o próximo passo é analisar a simetria da função de covariância, por analogia aos campos aleatórios espaciais, a função de covariância é simétrica se em qualquer direção de distâncias iguais de

espaço e tempo a correlação entre os atributos é a mesma. Sendo assim:

$$C(h,u) = C(h,-u) = C(-h,u) = C(-h,-u),$$

sendo $C(h,u) = Cov(Z(s_i,t_i), Z(s_j,t_j))$.

Definido o comportamento do campo aleatório, quanto a estacionariedade e simetria da função de correlação, agora há a necessidade de avaliar se a função de correlação adotada é separável ou não, ou seja, se na função de correlação é possível separar ou não os efeitos das correlações do tempo e espaço, devido a maior simplicidade de funções separáveis, essa abordagem é mais utilizada na prática. A separabilidade da função de correlação pode ser assegurada pelas regras de aditividade ou multiplicidade, propriedades análogas às propriedades das funções de covariância dos campos aleatórios espaciais. As propriedades de aditividade e multiplicidade para campos aleatórios espaço-temporais são dadas por:

$$Cov(Z(s_i,t_i), Z(s_j,t_j)) = Cov(Z(s_i,s_j)) + Cov(Z(t_i,t_j)),$$

ou

$$Cov(Z(s_i,t_i), Z(s_j,t_j)) = Cov(Z(s_i,s_j)) \cdot Cov(Z(t_i,t_j)).$$

Uma observação importante é que se a função de correlação é separável, a simetria da função de covariância está assegurada. Sendo assim, se pode utilizar funções de covariância conhecidamente positivas definidas, como por exemplo a função exponencial potência.

No entanto, na realidade não é compatível com a realidade trabalhar com estruturas de covariâncias que sejam separáveis, pois essa abordagem desconsidera a possível e muito freqüente interação entre tempo e espaço. Sendo assim, há a necessidade de levar em consideração modelos de covariância não separáveis. Duas famílias de funções covariância estacionária e não separáveis muito utilizadas são as propostas por Cressie-Huang (1999) e por Gneiting (2002).

Representação de Cressie-Huang ,

$$C(h,u) \propto \exp(-\|h\|^2/u^2 + c_0)\exp(-\delta u^2)/(u^2+c_0)^{d/2}, \tag{5}$$

a função (5) é estacionária, não separável e válida, positiva definida.

Representação de Gneiting

$$C(h,u) = \sigma 2 \cdot \phi(\|h\|^2/\psi(|u|^2))/(\psi(u)^2)^{d/2}, \tag{6}$$

a função (6) considera qualquer função monótona ϕ e qualquer função positiva ψ, ambas para $x \geq 0$.

3.2.1 Estimação dos parâmetros

As mesmas técnicas de estimação dos parâmetros em campos aleatórios espaciais podem ser utilizadas em campos aleatórios espaço-temporais, ou seja, pode-se estabelecer estimativas aos parâmetros através do variograma ou da função de verossimilhança. Na literatura de geoestatística, normalmente, é apresentado o semivariograma e não o variograma, segue a definição do semivariograma sob estacionariedade:

$$Var(Z(s,t) - Z(s+h,t+u)) = 2(Var(Z(s,t)) - C(h,u)) = 2\gamma(h,u). \tag{7}$$

Devido a relação, definida em (7), entre o semivariograma e a função de correlação, pode-se trabalhar com os dois para fazer estimação dos parâmetros, no entanto, os valores reais de variância e covariância são desconhecidos, daí a necessidade de analisar o semivariograma empírico:

$$\widehat{\gamma}(h) = (\sum_{|N(h,u)|} (Z(s_i,t_i) - Z(s_j,t_j))^2)/2|N(h,u)|, \tag{8}$$

sendo $|N(h,u)|$ o número de pontos dentro da distância h para cada distância u de tempo.

No entanto, fazer estimação através de (8) não é muito seguro, uma vez que os valores dependem muito da amostra coletada, se existirem poucos pontos, podem ocorrer valores de semivariogramas calculados com poucas observações dentro da distância h, ou ainda o semivariograma empírico pode ser muito distinto do semivariograma dependendo do acaso amostral. Sendo assim, os estimadores de máxima verossimilhança e de máxima verossimilhança restrita, que se baseiam na distribuição de probabilidade e nos valores dos dados para calcular as estimativas dos parâmetros, são os mais indicados na literatura de geoestatística, por outro lado, novamente, este tipo de estimação demanda muito tempo computacional, devido à dimensão da matriz de variância e covariância. E assim como em

campos aleatórios gaussianos, se for assumida a normalidade do campo aleatório espaço-temporal, a função de verossimilhança é igual a densidade de uma normal multivariada, no entanto agora pensa-se nos parâmetros como desconhecidos.

3.2.2 Predição

Em pesquisas com campos aleatórios espaço-temporais obviamente não é possível coletar informações de todo espaço e todo tempo de interesse, no entanto, na maioria dos casos, o interesse final é ter informação sobre todo espaço e em qualquer tempo, sejam eles amostrados ou não, sendo assim, técnicas de predição são utilizadas similarmente as ferramentas utilizadas em campos aleatórios espaciais, ou seja, os valores preditos aos locais e tempos não amostrados, são calculados através da média estimada para o processo com uma correção baseada nas estimativas de variabilidade e correlação. Cabe ressaltar que é confiável interpolar o campo aleatório espaço-temporal, no entanto, extrapolar a predição pode gerar problemas para as interpretações, por exemplo, não é muito confiável prever o valor do atributo sob estudo em uma determinada posição do espaço de interesse em um tempo futuro à pesquisa.

3.3 Campos aleatórios multivariados

O campo aleatório possui, agora, mais de uma variável resposta de interesse, ou seja, existem vetores $Y(s_i)$ de dimensão p x 1 , observados na posição espacial s_i do espaço de interesse d-dimensional G, a dimensão p é igual ao número de variáveis resposta sob estudo. Sendo assim, o modelo geoestatístico deve capturar a correlação entre os valores dentro de cada atributo e a correlação entre os atributos.

Um campo aleatório multivariado Z possui uma matriz de covariância Σ_Z de dimensão np x np, sendo n é o número de posições amostradas no espaço G, uma vez que, Σ_Z deve ser positiva definida, o maior problema nesse tipo de estudo é encontrar uma função de covariâncias $(C(s, s'))_{\iota,\nu} = Cov(Z(s)_\iota, Z(s')_\nu)$ válida, sendo $Z(s)_\iota$ o valor do campo aleatório para a variável ι na posição s e $Z(s')_\nu$ o valor observado do campo aleatório para a variável ν na posição s', para todo ι, ν, s e $s_\prime$.

Na literatura três abordagens são utilizadas para encontrar estruturas de covariâncias cruzadas válidas para modelar campos aleatórios multivariados, são elas modelos

separáveis, convolução de Kernel e convolução de modelos de covariância, a utilização dessas técnicas depende das configurações associadas ao campo aleatório, quanto a estacionariedade e isotropia.

A abordagem mais utilizada na literatura para encontrar funções de covariâncias válidas, devido a sua maior simplicidade, é de modelos separáveis, que possui, de forma geral, a seguinte expressão para a matriz de covariância:

$$\Sigma_Z = C(s, s') = \rho_{s,s'} T,$$

sendo T uma matriz positiva definida de dimensão p x p e ρ uma função de correlação univariada válida. Segundo Mardia e Gooddall (1993) e Banerjee e Gelfand (2002) a expressão definida em (9) é uma matriz de covariância válida.

Como os modelos separáveis consideram estacionariedade do processo estocástico, tem-se que as associações entre os valores dentro de cada variável e entre as variáveis são atenuadas conforme a aumenta distância entre os pontos espaciais amostrados. Se, além de estacionário, o processo for isotrópico, a matriz $\Sigma_Z = R \otimes T$, sendo R uma matriz n x n com $(R)_{i,j} = \rho(s, s')$.

Detalhes sobre as técnicas de convolução para encontrar funções de correlação válidas são apresentados por Gelfand et al. (2005).

Com relação a campos aleatórios que não são estacionários, Brown et al. (1994) e Sun et al. (1998) apresentam especificações para esse tipo de configuração de funções aleatórias multivariadas.

3.3.1 Modelos de co-regionalização

Como visto anteriormente, existem vetores $Y(s_i)$ de dimensão p x 1 , observados na posição espacial s_i do espaço de interesse d-dimensional G, a dimensão p é igual ao número de variáveis resposta sob estudo. Para modelar Gelfand et al. (2005) propuseram o seguinte modelo:

$$Y(s) = X(s)\beta + Z(s) + \epsilon(s),$$

$X(s)$ uma matriz np x q contento q possíveis covariáveis, β um vetor q x 1 de parâmetros associados as covariáveis, $Z(s)$ o campo aleatório , um processo latente, que existe mas não

pode ser medido, esse fator do modelo é responsável por capturar a estrutura de covariância cruzada entre as respostas e $\epsilon(s)$ é o ruído brando associado ao processo de amostragem, geralmente esse erro aleatório é assumido como normalmente distribuído com vetor de médias nulo e matriz de covariância p x p.

A idéia dos modelos de co-regionalização, assumindo alguma estacionariedade do campo aleatório, é decompor o termo $Z(s)$ do modelo estatístico acima da seguinte maneira:

$$Z(s) = A\omega(s),$$

sendo A uma matriz p x p de posto completo, que captura as variâncias associadas às variáveis aleatórias em cada posição espacial e $\omega(s)$ são processos espaciais independentes e identicamente distribuídos, e cada $\omega_j(s)$ possui média zero, e sob estacionariedade, variância 1 e função de correlação $\rho(s-s')$, assumindo-se que essa correlação só depende das distâncias entre as observações.

Sendo assim, E($\omega(s)$) é vetor nulo e esse processo latente possui matriz de covariâncias cruzadas $\Sigma_{Z(s)} = C(s-s') = \rho(s-s')AA'$, ou seja, considerando que $AA' = T$, existe equivalência com modelos separáveis e a função de covariância é válida.

Com a estrutura do processo latente definida, o próximo passo é pensar na modelagem de toda a estrutura do modelo estatístico proposto por Gelfand et al. (2005), ou seja, é necessário atribuir distribuições ao ruído branco e às variáveis do campo aleatório. Devido as propriedades da distribuição gaussiana, ela é a mais utilizada na literatura.

Assumindo que o ruído branco possui distribuição gaussiana multivariada com vetor de médias nulo e matriz de covariância D de dimensão p x p e com todos os elementos de covariância nulos, ou seja, uma matriz diagonal, além disso, se o campo aleatório possui distribuição gaussiana multivariada com vetor de médias nulo de dimensão np x 1 e matriz de covariância cruzada $\sum_{j=1}^{p} R_j \otimes T_j$, sendo R_j uma matriz n x n $(R_j)_{ii'} = \rho_j(s_i - s_{i'})$, então a distribuição de probabilidade de $Y(s)$ dado os parâmetros de média, de variância e covariância e de erro aleatório é definida como:

$$f(Y(s)|\theta) = N(\mu, \sum_{j=1}^{p} R_j \otimes T_j + I_{nxn} \otimes D), \tag{9}$$

sendo θ o vetor de todos os parâmetros associados ao modelo.

Com a distribuição conjunta de $Y(s)$ estabelecida em (9), pode-se pensar agora em utilizar os estimadores de máxima verossimilhança e máxima verossimilhança restrita para fazer a estimação dos parâmetros, o que pode ser computacionalmente inviável, uma vez que, a dimensão da matriz de covariância do campo aleatório cresce exponencialmente com o aumento do número de variáveis respostas. Logo, métodos Bayesianos e métodos computacionais intensivos são muito utilizados nesse tipo de abordagem, onde utiliza-se alguma informação a priori sobre a distribuição dos parâmetros.

Uma abordagem concorrente a da distribuição conjunta dos valores observados do campo aleatório multivariado é a utilização de distribuições condicionadas entre as respostas. Para simplificar, considerando que o campo aleatório possui duas variáveis de interesse, deve-se pensar na seguinte modelagem aos dados:

$$Y_1(s) = X^T(s)\beta_1 + \sigma_1\omega_1(s) \tag{10}$$

$$Y_2(s)|Y_1(s) = X^T(s)\beta_2 + \alpha^{2|1}Y_1(s) + \sigma_2\omega_2(s) + \epsilon(s) \tag{11}$$

sendo que a primeira variável não possui ruído branco associado e explica uma parte da variabilidade da segunda variável, fato que explica a entrada da primeira variável como covariável da segunda.

Com a estrutura definida em (10) e (11), tem-se a seguinte função de verossimilhança:

$$L(\theta|Y(s)) = f(Y_1(s)|\theta_1)f(Y_2(s)|Y_1(s), \theta_2),$$

sendo θ_1 e θ_1 os vetores de parâmetros associados aos modelo para $Y_1(s)$ e $Y_2(s)|Y_1(s)$, respectivamente, e a união dos dois vetores é o vetor θ que contém todos os parâmetros do modelo.

A vantagem da abordagem condicional é que se for assumido que a distribuição a priori para θ é igual ao produto das distribuições a priori de θ_1 e θ_2, ou seja, independentes, o condicionamento resulta na fatorização de dois modelos e assim, cada um deles pode ser ajustado em separado.

Todos os resultados apresentados até aqui para campos aleatórios multivariados foram específicos para estruturas de covariâncias estacionárias, no entanto, na prática

podem ocorrer problemas em essa suposição não é válida. Sendo assim, quando um processo espacial multivariado não possui estacionariedade da matriz de covariância, deve-se pensar na seguinte modelagem para o campo aleatório:

$$Z(s) = A(s)\omega(s),$$

e similarmente a processos com função de covariância estacionária, pode-se pensar na notação $T(s) = A(s)A(s)T$ e assim a matriz de covariância possui as seguintes entradas:

$$C(s, s') = \sum_j \rho_j(s - s')a_j a_j^T(s'), \tag{12}$$

sendo $a_j(s)$ a j-ésima coluna de $A(s)$, $T_j(s) = a_j(s)a_j^T(s)$ e assima $\sum_j T_j(s) = T(s)$.

Com as definições estabelecidas em (12), novamente o problema é modelar uma matriz de covariância para o campo aleatório válida, as mesmas técnicas com pressupostos de estacionariedade podem ser utilizadas, e agora pode-se modelar esse tipo de problema utilizando a distribuição de Wishart, Gelfand et al. (2004) apresentam maiores detalhes sobre esse tipo de ferramenta.

3.4 Cópulas

Segundo Gomes (2007) cópulas são funções que fornecem um meio de relacionar funções de distribuições multivariadas com funções de distribuição marginais e apresentam enorme flexibilidade na escolha das marginais, flexibilidade que não é muito utilizado em estudos geoestatisticos multivariados clássicos, pois pode gerar em problemas para determinar as distribuições conjuntas associadas ao processo espacial.

Seja Y_j uma variável uma variável aleatória com função de distribuição marginal contínua F_j, $j = 1, 2, ..., k$, a função de distribuição conjunta será da forma:

$$F(y_1, y_2, ..., y_k) = C_\alpha(F_1(y_1), F_2(y_2), ..., F_k(y_k)),$$

em que $C_\alpha(F_1(y_1), F_2(y_2), ..., F_k(y_k))$ é a cópula do vetor aleatório $(Y_1, Y_2, ..., Y_k)$, tendo como resultado da inferência estatística que $F_j(Y_j) \sim U(0, 1)$ para todo j.

Utilizando a abordagem de cópulas, o parâmetro ou conjunto de parâmetros α descreve completamente a dependência entre as variáveis aleatória. Na literatura, as cópulas

Archimedianas uniparamétricas são muito utilizadas em modelos bivariados, dois tipos de cópulas Archimedianas que são muito utilizadas é a de Frank e a de Clayton, Gomes (2007) apresenta maiores detalhes sobre as formas dessas cópulas. Uma observação importante é que cópulas só devem ser utilizadas se existir uma forte suspeita de que as variáveis sob estudo são fortemente correlacionadas.

4 MATERIAL E MÉTODOS

A Comissão Internacional para a Conservação dos Atuns do Atlântico - ICCAT (2007) disponibiliza em seu site (www.iccat.int) conjuntos de dados sobre captura de algumas espécies de atuns (em kg) e esforço desprendido para tal captura, todas as observações possuem informações de latitude e longitude onde ocorreu a pescaria, o que já caracteriza um possível problema geoestatístico, no entanto, também existem informações temporais das pescarias, ou seja, o problema agora passa a ser modelagem espaço-temporal, com resposta bivariada, captura e esforço. A importância desse tipo de estudo é fornecer informações refinadas sobre o estoque de atuns no Atlântico Sul, o que pode nortear tomadas de decisões governamentais quanto a políticas de preservação da espécie.

No entanto antes da aplicação das técnicas estatísticas, é necessário estudar o problema das técnicas de geoestatística possuírem suas estruturas de correlação baseadas em distância, geralmente euclidiana, entre as observações. E como os dados possuem coordenadas em latitude e longitude em uma região muito grande do planeta, as distâncias euclidianas não irão funcionar bem, pois em regiões próximas à linha do equador os valores de latitude e longitude se aproximam, possuem escalas parecidas, o que vai se distorcendo a medida que coordenadas vão se aproximando dos pólos. Devido a esse problema é necessário fazer alguma transformação nas coordenadas, uma muito utilizada na literatura é a projeção Universal Transversa de Mercator (UTM), para fazer tal transformação nas coordenadas serão utilizados softwares de Sistemas de Informações Georeferenciadas (SIG).

Após contornar os problemas de escala das coordenadas, a proposta é estabelecer um modelo espaço-temporal para captura e esforço conjuntamente, para tal serão utilizadas as abordagens multivariadas das distribuições de probabilidade conjuntas e das cópulas.

No entanto, devido à complexidade dos modelos que serão propostos e ao número de parâmetros envolvidos nas estruturas de correlação, técnicas clássicas de estimação e inferência não são trivialmente desenvolvidas, sendo assim, outras abordagens, com usos mais recentes, serão consideradas: a abordagem Bayesiana e métodos computacionais intensivos.

Também será feito um estudo de simulação de dados para analisar detalhada-

mente o comportamento dos modelos propostos e de seus pressupostos com diferentes configurações de distribuições de probabilidade e diferentes configurações paramétricas.

Parte dos resultados dessa dissertação serão gerados com o auxílio computacional, o ambiente R de programação e seus pacotes serão o mais utilizados. Ainda não existe um pacote específico para estudos de modelos espaço-temporais com resposta bivariada, sendo assim, será necessária a exploração de funções já implementadas para modelar os problemas levantados, sendo que a possível criação de funções no ambiente R pode gerar mais áreas de pesquisas.

REFERÊNCIAS

BANERJEE, S.; GELFAND, A.E. Predict, Interpolation and regression for spatial misaligned data points. **Sankhya**, v.64, p.227-245, 2002.

BROWN, P.J.; LE, N.D.; ZIDEK, J.V. Multivariate spatial interpolation and exposure to air pollutants. **The Canadian Journal of Statistics**, v.22, p.489-509, 1994.

CRESSIE N.; HUANG, H-C. Classes of Non-Separable, Spatio-temporal stationary covariance functions. **Journal of the American Statistical Association**, Alexandria, v.94, p.1330-1340, 1999.

DIGGLE, P.J.; RIBEIRO Jr., P.J. **Model-Based geostatistics**. New York: Springer, 2006. 230p.

ELMATZOGLOU, I. **Spatio-temporal geostatistical models, with an aplication in fish stock**, 2006. 53 p. Submited for the degree of (Master in statistics) - Lancaster University, Lancaster, 2006.

FERNANDES, M.V.M. **Modelos para Processos Espaço-Temporais Inflacionados de Zeros**, 2006. 128 p. Dissertação (Mestrado em Estatística) - Instituto de Matemática da Univeridade Federal do Rio de Janeiro, 2006.

FUENTES, M.; SMITH, R.L. **A new class of stationary spatial models**. North Caroline: Department of Statistics, North Caroline State University, 2001. Technical (Report, 2534).

GELFAND, A.E; SCHMIDT, A.M.; BANERJEE S.; SIRMANS, C.F. Nonstationary multivariate process modeling through spatially varying coregionalization. **Sociedad Española de Estadística e Investigación Operativa - Test**, v.13, p.263-312, 2005.

GOMES, E.M.C. **Modelos de regressão com resposta bivariada através de cópulas: Análise de sensibilidade e resíduo**, 2007. 101 p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura ″Luiz de Queiroz" da

Univeridade de São Paulo, 2007.

GNEITING, T. Nonseparable, Stationary Covariance Functions for Space-Time Data. **Journal of the American Statistical Association**, Alexandria, v.97, p.590-600, 2002.

HIGDON, D. **Quantitative methods for current environmental issues**, Chichester: Wiley, 2002. 185 p.

LE, D.N.; ZIDEK, J.V. **Statistical analysis of environmental space-time processes.** New York: Springer, 2006. 327p.

MARDIA, K.V.; GOODALL, C.R. Spatial-temporal analysis of multivariate environmental monitoring data. In: G. P. PATIL and C. R. Rao, eds., **Multivariate Environmental Statistics**, p.347-386, 1993.

MATÉRN, B. **Spatial variation**. Verlag, Berlin: Spinger, 1986. 365 p.

SAMPSON P.D.; GUTTORP, P. Nonparametric estimation of nonstationary spatial covariance structure. **Journal of American Statistical Association**, Alexandria, v.87, p.108-119, 1992.

SCHABENBERGER, O.; GOTWAY, C.A. **Statistical methods for spatial data analysis**, Boca Raton: Chapman and Hall / CRC, 2005. 488p.

SCHMIDT, A.M.; O'HAGAN, A. Bayesian inference for nonstationary spatial covariance structure via spatial deformations, **Journal of Royal Statistical Society**, Oxford: v.65, p.743-758, 2003.

SCHMIDT, A.M.; SANSÓ, B. Modelagem bayesiana da estrutura de covariância de processos espaciais e espaço-temporais. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 14, 2006. Caxambú, **Minicurso**, São Paulo: Associação Brasileira de Estatística, 2006. 151 p.

SILVA, A.S. **Modelos gaussianos geoestatísticos espaço-temporais e aplicações**,

2006. 70 p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura ″Luiz de Queiroz” da Univeridade de São Paulo, 2006.

SUN, W.; LE, N.D.; ZIDEK, J.V.; BURNETT, R. Assessment of a bayesian multivariate interpolation approach for health impact studies. **Environmetrics**, Washington, v.9, p.565-586, 1998.