UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS

BRUNA DE CASTRO DIAS BICALHO

Modelos Espaço–Temporais: Estudo de Casos

Belo Horizonte
2008

BRUNA DE CASTRO DIAS BICALHO

Modelos Espaço–Temporais: Estudo de Casos

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Estatística.

Orientadora: Profa. Dra. Sueli Aparecida Mingoti

Belo Horizonte

2008

DEDICATÓRIA

Dedico esta dissertação às pessoas da minha família, guias nas mais difíceis trilhas, mestres no ensinamento, real exemplo de dedicação, companheirismo e ética.

E se em algum momento da vida, algum deles por fortuito precisar predizer observações em locais e/ou tempos não amostrados, ofereço, os resultados desta dissertação, como forma de gratificação.

AGRADECIMENTOS

À Alcione, Adriana, Ana e Lilian. Primeiros ensaios de convívio, minhas melhores referências;

Ao Gift, Isabel, Gabriel e Bruninho pela companhia constante;

Ao Almir pela paciência e generosidade;

Às amigas, Edimeire e Flávia, com as quais eu aprendi e compartilhei as alegrias e as angústias do fazer mestrado;

À Profa. Dra. Sueli Aparecida Mingoti pela orientação e confiança depositada, e que por meio de constantes desafios soube me encorajar a explorar as minhas potencialidades;

Aos professores Dra. Ela Mercedes Medrano de Toscano, Dr. Luiz Henrique Duczmal e Dr. Sabino José Ferreira Neto pelas sugestões e discussões construtivas que se delinearam durante a qualificação;

Ao professor Dr. Paulo Justiniano Ribeiro Jr. pelo intercâmbio de idéias e incentivos, que foram de fundamental importância para o desenvolvimento do meu projeto;

Aos colegas Alexandre Sousa da Silva e Elias Teixeira Krainski pela inestimável ajuda;

Ao Consórcio de Informações Sociais – CIS e ao departamento de informática do SUS – DATASUS, órgão da Secretaria Executiva do Ministério da Saúde, pela disponibilidade dos dados utilizados neste trabalho;

À Dolorice Moreti por ter gentilmente cedido os dados de armazenagem de água no solo;

Ao Marcos e a Renata Mendonça do Centro de Previsão de Tempo e Estudos Climáticos – CPTEC;

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES pela concessão da bolsa de mestrado.

“All models are wrong, but some are useful”

George E. P. Box

RESUMO

BICALHO, B. D. C. D. Modelos Espaço–Temporais: Estudo de Casos. 2008. 174f. Dissertação (Mestrado) – Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

Os dados provenientes de diversas áreas tais como ciências ambientais, biologia, epidemiologia, agricultura, sociologia etc. são caracterizados pela variabilidade no espaço e no tempo. Os procedimentos de estatística, freqüentemente, não consideram a interação entre as dimensões espacial e temporal, e nos últimos anos tem ocorrido um aumento crescente no desenvolvimento de técnicas para a análise de processos desta natureza devido principalmente, a grande aplicabilidade dos modelos espaço-temporais. O objetivo da análise de processos espaço-temporais, na maioria dos casos, resume-se na predição de observações em localizações e/ou tempos não amostrados. Segundo Schabenberger e Gotway (2005) há uma carência de softwares que lidam com dados no espaço e no tempo conjuntamente e a maioria dos estudos da área utiliza-se de dados relacionados a fenômenos ambientais. Nesta dissertação implementamos computacionalmente alguns modelos muito citados na literatura como o de Høst et al. (1995), o de Kyriakidis e Journel (1999) e as funções de covariância propostas por Gneiting (2002). Além disso, implementamos o modelo espacial de séries temporais proposto por Niu et al. (2003) e uma nova estratégia de análise de dados combinando essa metodologia com a de Høst et al. (1995). Todos esses modelos foram ajustados a dados reais provenientes de áreas distintas sendo a qualidade do ajuste avaliada a partir das predições de observações no espaço e/ou no tempo. Para os conjuntos de dados considerados observamos que a adequação dos modelos está relacionada com a variação dos dados no espaço e no tempo, ou seja, os erros de predição são menores em localizações onde os vizinhos têm um comportamento espacialmente semelhante da característica de interesse e, além disso, nos pontos nos quais a série temporal não sofre variações bruscas ao longo do tempo.

Palavras-chave: Geoestatística, Modelos Espaço–Temporais, Predição, Estudo de Casos.

ABSTRACT

BICALHO, B. D. C. D. Space-Time Models: Study of Cases. 2008. 174f. Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

In several areas of study such as environmental sciences, biology, epidemiology, agriculture, sociology, etc. the sample data are characterized by spatial and temporal variation. Some of the statistical techniques do not consider the interaction between space and time dimensions. In the recent years many researches have been developed for the analysis of data of this nature. The main objective of the analysis of space-time processes, in the majority of the cases, is to predict observations at localizations and/or times not observed or sampled. In this dissertation some space-time models were implemented computationally and compared using some data sets. The models discussed in this dissertation are: Host et al. (1995) with some modifications proposed by Kyriakidis and Journel (1999), the covariance functions proposed by Gneiting (2002) and the space-time series model proposed by Niu et al. (2003) with some practical modifications. All these models were adjusted in three real data sets and their performance were compared. A practical strategy to analyze space-time data, which is a combination of the Host et al. (1995) and Niu et al. (2003) models, was also proposed. In this dissertation we also discussed some computation aspects of the implementation of all models. Gneiting (2002) functions were adjusted in RandomFields package from software R.

Keywords: Geostatistics, Space-Time Models, Prediction, Study of Cases.

SUMÁRIO

Capítulo 1

Introdução

As observações tomadas em diferentes posições no espaço e tempos/momentos são provenientes de processos espaço-temporais. Segundo Schmidt e Sansó (2006), a modelagem deste tipo de processo é uma área recente na estatística que vem sendo pesquisada nos últimos 15 anos e que se encontra em pleno desenvolvimento. Isto se deve principalmente ao avanço de recursos computacionais, tanto no que se refere à capacidade de armazenagem de enormes volumes de dados quanto no processamento dos computadores.

Outro fator de grande importância é a numerosa aplicabilidade dos modelos espaço-temporais em diversas ciências, especialmente naquelas relacionadas a fenômenos ambientais, tais como meteorologia, hidrologia e climatologia. Haslett e Raftery (1989), De Luna e Genton (2005), Stein (2005) e Porcu et al. (2007) entre outros estudam a velocidade do vento no período de 1961-1978 em 12 estações meteorológicas na Irlanda; Cressie e Huang (1999) também estudam a velocidade do vento em uma região tropical a oeste do oceano Pacífico; Handcock e Wallis (1994) analisam a temperatura média de uma região do norte dos Estados Unidos; Paez e Gamerman (2005) estudam a poluição atmosférica no Rio de Janeiro avaliando as concentrações de PM_{10} (partículas inaláveis com diâmetro menor que $10\,\mu g/m^2$); Huerta, Sansó e Stroud (2004) avaliam a qualidade do ar a partir dos dados da concentração de ozônio na Cidade do México; De Iaco, Myers e Posa (2002, 2003) trabalham com dados da concentração média por hora de NO_2 e CO (mg/m^3) em 18 estações de medição em Milão; Brown et al. (2001) avaliam a intensidade da chuva em Lancashire, Inglaterra; Niu et al. (2003) estudam a densidade do ar no hemisfério setentrional com o objetivo de predizer o clima a médio e em longo prazo, entre outros.

As preocupações com os problemas ambientais emergentes no século XXI, como o aquecimento global, também têm motivado o desenvolvimento de ferramentas para a análise de dados espaço-temporais, pois existe um interesse generalizado na predição da temperatura global média dentro de alguns anos. Este interesse está relacionado às conseqüências devastadoras deste fenômeno sobre a saúde humana, a economia e o meio ambiente, tais como a ameaça de submersão de cidades litorâneas devido ao aumento do nível dos oceanos com o derretimento das calotas polares, o crescimento e o surgimento de desertos resultantes

de desequilíbrios do ecossistema e o aumento de furacões, tufões e ciclones devido a maior evaporação das águas dos oceanos.

Os modelos espaço-temporais procuram descrever probabilisticamente a variabilidade dos dados no espaço e no tempo. O interesse primordial da análise espaço-temporal consiste, na maioria das vezes, na predição de observações em locais e/ou tempos não amostrados. Desta forma observamos três casos distintos de análise: no primeiro caso o objetivo do estudo é a interpolação espacial em tempos observados na amostra; no segundo caso o propósito da análise é a previsão temporal em localizações amostradas; e o terceiro caso é referente a situações onde as predições são realizadas em locais e tempos não observados na amostra.

Os procedimentos de estatística, freqüentemente, não são suficientes para descrever os processos espaço–temporais, pois não consideram a interação entre o espaço e o tempo, ou seja, estes procedimentos não conseguem captar a variabilidade nas dimensões espaço e tempo conjuntamente.

Schabenberger e Gotway (2005) enumeram três procedimentos para modelar processos espaço-temporais. Os dois primeiros consideram a análise separada do espaço e do tempo, ou seja, isola-se a parte espacial ou a parte temporal e aplicam-se técnicas estatísticas padrões para o tipo de processo resultante. Estes dois procedimentos são úteis apenas como ferramentas de análise descritiva de dados, pois a análise separada do tempo (ou espaço) não permite predições em tempos futuros (ou em novas localizações).

O terceiro procedimento, citado por Schabenberger e Gotway (2005), realiza a análise conjunta destas duas dimensões utilizando métodos para campos aleatórios definidos em R^{d+1}, onde d é a dimensão espacial. O espaço e o tempo são dimensões que não podem ser diretamente comparáveis, pois no espaço não existe a idéia de ordenação (passado, presente e futuro) que é bem definida no tempo, e em contrapartida, no espaço existem os conceitos de isotropia e anisotropia que não têm sentido no tempo. Desta forma as funções de covariância devem refletir esta diferença (física) de grandezas a partir das matrizes de anisotropia.

A dificuldade em modelar processos correlacionados no espaço e no tempo é motivada, em parte, pela necessidade de modelos de covariância não-separáveis, pois estes modelos requerem que as funções de covariância sejam válidas atendendo a condição de serem positivas definidas (MA, 2003).

Gneiting (2002) propõe um procedimento para a obtenção de funções de covariância espaço-temporal separáveis e não-separáveis válidas a partir da combinação de funções completamente monótonas e funções positivas com derivadas completamente monótonas sendo que este método não necessita de operações no domínio espectral.

Høst et al. (1995) propõem um modelo espaço-temporal que é uma extensão de um modelo geoestatístico e que inclui a componente temporal, ou seja, incorpora-se na predição espacial a informação dos vizinhos ao longo do tempo, e conseqüentemente este modelo é incapaz de predizer observações em tempos não amostrados.

Kyriakidis e Journel (1999), em uma revisão abrangente sobre processos espaço-temporais, sugerem um procedimento alternativo para estimar as componentes do modelo proposto por Høst et al. (1995).

Niu et al. (2003) estudam uma classe de modelos denominada modelos espaço-temporais autoregressivos e de médias móveis. Os autores investigam os modelos temporais ARMA de Box e Jenkins e adicionam uma componente espacial na modelagem. Essa componente traz informações importantes a respeito da configuração espacial dos pontos e teoricamente esta modelagem deve fornecer melhor ajuste aos dados e resultar em menores erros de predições se comparada com os modelos puramente temporais. Estes modelos não fazem a interpolação espacial, i.e., somente conseguem predizer observações futuras nas localidades amostradas.

As metodologias de Høst et al. (1995) e de Kyriakidis e Journel (1999) embora muito mencionadas na literatura, não aparecem implementadas em trabalhos práticos publicados.

De acordo com Schabenberger e Gotway (2005) há uma carência de *softwares* comerciais que lidam com dados no espaço e no tempo conjuntamente. O *software* R (2006) apresenta um pacote chamado *RandomFields* (2006) que permite a análise e a simulação de dados espaciais e espaço–temporais, porém segundo Silva (2006, p. 30) “a modelagem espaço-temporal no pacote “RandomFields” está em desenvolvimento e por conta disso seus procedimentos ainda não são exaustivos e claramente documentados e existem poucos trabalhos que se utilizam deste pacote”.

A maioria dos trabalhos da área de processos espaço-temporais analisa dados relacionados a fenômenos ambientais como mencionamos previamente e, portanto também existe uma necessidade de avaliar o comportamento destes modelos propostos na literatura aplicados (ou ajustados) a dados provenientes de outras ciências tais como: biologia, epidemiologia e sociologia.

A seção seguinte apresenta os objetivos desta dissertação.

1.1 Objetivos

Diante da necessidade de estudos e programas computacionais para a análise de dados no espaço e no tempo conjuntamente, conforme a exposição feita previamente, esta dissertação teve como objetivos gerais:

- Fazer uma revisão da literatura sobre os modelos mais atuais da área de processos espaço–temporais;
- Implementar computacionalmente[1] o modelo espaço-temporal proposto por Høst et al. (1995) e a proposta de estimação das componentes deste modelo sugerida por Kyriakidis e Journel (1999);
- Implementar computacionalmente o modelo proposto por Niu et al. (2003) para a análise de dados reais;
- Ajustar modelos usando funções de covariância espaço-temporal propostas por Gneiting (2002) utilizando os recursos do pacote *RandomFields*;
- Aliar as idéias apresentadas no artigo de Niu et al. (2003) e o modelo proposto por Høst et al. (1995) na tentativa de construir um modelo capaz de predizer observações em locais e tempos não amostrados e que seria uma alternativa prática aos modelos de Gneiting (2002);
- Avaliar estes modelos utilizando dados reais provenientes de áreas distintas: ciências ambientais, epidemiologia e sociologia/psicologia, sendo que os dados procedentes destas duas últimas áreas citadas, em geral, apresentam maiores variações no espaço e no tempo se comparados com os dados relacionados a fenômenos ambientais;
- Comparar o ajuste destes modelos em três situações distintas de acordo com o tipo de predição: interpolação espacial em tempos observados na amostra, previsão temporal em localizações amostradas e predição em locais e tempos não observados na amostra;
- Introduzir uma estratégia para a análise de dados espaço-temporais.

1.2 Organização da Dissertação

Esta dissertação está dividida em sete capítulos. No Capítulo 2 fazemos uma revisão dos principais conceitos da metodologia de geoestatística. Estes conceitos são de suma

[1] A implementação computacional dos modelos foi feita no *software* R (2006).

importância para o entendimento de processos espaço-temporais, visto que estes processos são baseados em processos puramente espaciais.

No Capítulo 3 é apresentada a metodologia espaço-temporal onde discutimos os tipos de representações para caracterizar processos desta natureza e em especial, a função de covariância. Mostramos a família de funções de covariância propostas por Gneiting (2002), o modelo geoestatístico proposto por Høst et al. (1995), o modelo de séries temporais de Niu et al. (2003), e também sugerimos algumas modificações nos modelos de Høst et al. (1995) e de Niu et al. (2003). Além disso, propomos uma estratégia de modelagem que combina os modelos de Høst et al. (1995) e de Niu et al. (2003) em duas etapas distintas de análise.

Os Capítulos 4, 5 e 6 mostram os resultados das aplicações dos modelos apresentados no Capítulo 3 a dados reais referentes respectivamente a: taxa de criminalidade violenta nas regiões administrativas do estado de Minas Gerais (MG), a armazenagem de água em um solo cultivado com citros e a incidência de AIDS nas microrregiões de MG.

Finalmente, no Capítulo 7, fazemos as considerações finais.

Capítulo 2

Metodologia de Geoestatística: Modelos Espaciais

No capítulo 2 apresenta-se uma revisão dos principais conceitos da metodologia de geoestatística onde definimos: estatística espacial, geoestatística, função aleatória, estacionariedade, variograma, krigagem e validação cruzada.

2.1 Estatística Espacial e Tipo de Dados

A estatística espacial considera a localização dos dados do evento de interesse no estudo, ou seja, o local onde a informação é coletada é relevante e deve ser considerado na análise.

Segundo Cressie (1993), a estatística espacial divide-se em três grandes áreas conforme o tipo de dado analisado: processos pontuais ou eventos, geoestatística ou dados espacialmente contínuos e dados de área ou lattice. Bailey e Gatrell (1995) consideram ainda uma quarta classe que é definida como dados de interação.

Os processos pontuais não consideram a realização de uma variável aleatória no local amostrado, isto é, a própria localização do fenômeno é o evento no qual estamos interessados. Assim, a variável indicativa, ocorrência ou não do evento, define o processo pontual. O objetivo básico é avaliar a distribuição espacial dos pontos verificando se estes exibem um comportamento sistemático, apresentando-se como aglomerados (“clusters”) ou regularmente distribuídos, ou se é aleatório. Exemplos incluem: ocorrências de doenças, localização de crimes, certos tipos de árvores em uma floresta e epicentros de terremotos. Diggle (2003) descreve detalhadamente os processos pontuais.

De acordo com Diggle e Ribeiro Júnior (2007) o termo *geoestatística* tem sua origem na França, na escola de Fontainebleau, para tratar de problemas de predição associados à mineração. Os dados de geoestatística variam continuamente dentro da área de estudo e as localizações, onde são feitas as medidas, são fixas e podem estar regularmente ou irregularmente distribuídas. Em geral, quando trabalhamos com este tipo de dado, o interesse é predizer valores em locais não amostrados e/ou recuperar a superfície. São exemplos: concentração de poluentes, medidas de chuva, temperatura e determinação de propriedades do

solo. As técnicas de geoestatística podem ser aplicadas em problemas diversos, além daqueles relacionados a fenômenos naturais. Como exemplos, Mingoti et al. (2006) e Pantuzzo e Mingoti (1998) usam a metodologia de geoestatística para predizer o número total de casos diagnosticados de AIDS em alguns municípios do estado de Minas Gerais; Cressie e Chan (1989) avaliam as taxas de mortalidade infantil; Diggle e Ribeiro Júnior (2007) apresentam dois estudos de casos, o primeiro trata do nível de radiação na ilha de Rongelap e o segundo avalia a prevalência de malária em crianças de Gâmbia, oeste da África.

Os dados de área não apresentam uma variação contínua no espaço, o que os distingue da geoestatística, pois associamos ao atributo de interesse uma região dentro da área de estudo. Supomos então que a característica avaliada é homogênea dentro de cada região. Alguns exemplos são: nível de escolaridade médio por município e medidas sócio-econômicas por cidade. Banerjee et al. (2004) tratam destes dados usando uma abordagem bayesiana.

Nesta dissertação estamos trabalhando com a modelagem espaço–temporal de dados provenientes da segunda área, geoestatística. As próximas seções apresentam os principais conceitos desta metodologia.

2.2 Processo Estocástico

Seja $Z(s)$ a variável aleatória ou o atributo de interesse medido na localização $s \in D \subset \Re^d$, onde D é a região de estudo e $z(s)$ é uma realização da variável, isto é, o valor observado, sendo $d = 1,2,...$ a dimensão do campo aleatório. Geralmente estamos trabalhando em um espaço bidimensional ($d = 2$), por exemplo, quando consideramos as coordenadas latitude e longitude. O processo estocástico (ou função aleatória ou campo aleatório) é definido como uma família de variáveis aleatórias reais indexadas pela posição e que variam continuamente no espaço. Usualmente estas variáveis são dependentes e supomos que elas estão definidas num mesmo espaço de probabilidade (Ω, A, P).

O processo estocástico $Z = \{Z(s), s \in D \subset \Re^d\}$ é determinado pela suas distribuições finito-dimensionais (CRESSIE, 1993),

$$F_{s_1, s_2, ..., s_k}(z_1, z_2, ..., z_k) = P(Z(s_1) \leq z_1, Z(s_2) \leq z_2, ..., Z(s_k) \leq z_k), \quad \forall k \geq 1 \tag{2.1}$$

Isto significa que, devemos conhecer a distribuição conjunta de $\{Z(s_1), Z(s_2), ..., Z(s_k)\}$ para todo valor de *k* finito e qualquer configuração dos pontos $(s_1, s_2, ..., s_k)$.

Um processo estocástico é gaussiano se a distribuição conjunta de $\{Z(s_1), Z(s_2), ..., Z(s_k)\}$ é uma normal multivariada, para quaisquer escolhas de *k* e localizações $(s_1, s_2, ..., s_k)$. Isso implica que cada variável aleatória $Z(s_i), i = 1, 2, ..., k$, segue uma distribuição normal. Temos que a distribuição normal *k*-variada é completamente especificada quando conhecemos o vetor de médias e a matriz de covariâncias (SCHMIDT; SANSÓ, 2006).

Na prática, frequentemente, nós observamos apenas uma única realização da função aleatória. A proposta da geoestatística para inferir sobre a distribuição do processo e os seus momentos utiliza todos os pares de observações separados pelo vetor de distância *h* como sendo medidas repetidas da variável aleatória. Este procedimento é baseado na dependência espacial entre as observações, isto é, na idéia intuitiva de que a semelhança entre as medições é mais forte à medida que diminuímos a distância entre as localizações. Para proceder na análise dos dados espaciais a inferência estatística deve então considerar algumas hipóteses sobre a estabilidade do processo. Estas suposições são discutidas a seguir quando definimos os três tipos de estacionariedade: estrita, de segunda ordem e intrínseca.

2.3 Estacionariedade

Um processo estocástico $Z = \{Z(s), s \in D \subset \Re^d\}$ é dito ser estritamente estacionário quando as distribuições finito-dimensionais definidas em (2.1) permanecem as mesmas para qualquer translação determinada pelo vetor *h* e para todos os pontos $(s_1, s_2, ..., s_k)$, ou seja, (GOOVAERTS, 1997):

$$F_{s_{1+h}, s_{2+h}, ..., s_{k+h}}(z_1, z_2, ..., z_k) = F_{s_1, s_2, ..., s_k}(z_1, z_2, ..., z_k) \tag{2.2}$$

Neste caso assumimos que todos os momentos não variam quando transladamos a origem do sistema de coordenadas, onde coletamos as observações. Devido à limitação dos dados observacionais uma suposição mais fácil de ser verificada é a de estacionariedade fraca ou de segunda ordem, a qual requer que somente os dois primeiros momentos, média e covariância, sejam invariantes sob translações. Em outras palavras, o valor esperado de $Z(s)$ existe e tem que ser constante para todas as localizações s, e a função de covariância entre

duas variáveis, $Z(s)$ e $Z(s+h)$ separadas por um lag h, depende apenas do vetor de separação h (comprimento e direção) e não da localização s, isto é, (CRESSIE, 1993):

$$\begin{cases} E(Z(s)) = m(s) = m, \forall\, s \in D \\ Cov(h) = Cov(Z(s), Z(s+h)) = E[Z(s)Z(s+h)] - m^2, \forall\, s \in D \end{cases} \tag{2.3}$$

A variância é uma função da covariância quando consideramos a distância igual à zero, $Cov(0) = Var(Z(s)), \forall\, s \in D$, logo não precisamos fazer suposições sobre a variabilidade do processo.

Quando a função de covariância (2.3) depende apenas da magnitude de *h*, sendo a direção e o sentido do vetor irrelevante, temos um processo isotrópico. Caso contrário denomina-se um campo aleatório anisotrópico. Podemos ter a anisotropia geométrica, que é a mais simples, ou a anisotropia zonal. O leitor interessado em maiores detalhes sobre o tipo de anisotropia deve consultar Armstrong (1998).

Conforme Cressie (1993, p.53, tradução nossa), "a estacionariedade estrita implica na estacionariedade de segunda ordem [fraca] se o segundo momento de F é finito". As duas suposições são coincidentes quando temos um processo estocástico gaussiano, pois a estacionariedade do processo implica que a variância de $Z(s)$ é constante e é igual a $Var[Z(s)] = \sigma^2, \forall s \in D$. Como dissemos anteriormente, este modelo de probabilidade é bem definido pelas funções de média e de covariância. Assim, temos que $Cov(h) = \sigma^2 \rho(h)$ onde $h = \|s_i - s_j\|, s_i, s_j \in D$ e $\rho(\cdot)$ é a função de correlação dada por $\rho(h) = corr\{Z(s+h), Z(s)\}$.

Uma condição menos restritiva, conhecida como estacionariedade intrínseca (JOURNEL; HUIJBREGTS, 1978), requer que os incrementos $[Z(s+h) - Z(s)]$, ao invés das variáveis aleatórias $Z(s)$, sejam estacionários de segunda ordem. A estacionariedade dos incrementos, ou seja, das diferenças entre as variáveis aleatórias separadas pela distância h diz que:

$$\begin{cases} E[Z(s+h) - Z(s)] = 0 \\ Var[Z(s+h) - Z(s)] = 2\gamma(h) \end{cases} \tag{2.4}$$

sendo $2\gamma(h)$ o variograma[2] e $\gamma(h)$ o semi-variograma. Quando o processo além de ser intrínseco for estacionário de segunda ordem, o variograma pode também ser escrito como:

[2] Alguns autores consideram $\gamma(h)$ como sendo o variograma ao invés da função $2\gamma(h)$. Schabenberger e Gotway (2005, p. 135) discutem brevemente a origem e o emprego do termo "variograma".

$$2\gamma(h) = E\left\{\left[Z(s+h) - Z(s)\right]^2\right\} \tag{2.5}$$

pois $\left\{E\left[Z(s+h) - Z(s)\right]\right\}^2 = 0$, pela estacionariedade fraca. A função $\gamma(\cdot)$ será discutida detalhadamente na seção 2.4.

Se nenhuma das hipóteses de estacionariedade é verificada, dizemos que o campo aleatório é não-estacionário. Segundo Diggle e Ribeiro Júnior (2007) a não estacionariedade pode ser resultante de uma *tendência espacial* e/ou uma *tendência de superfície*, isto é, a média do processo varia de acordo com a localização. Neste caso, podemos assumir um modelo de regressão polinomial para a média e tomar como variáveis explicativas as coordenadas cartesianas (tendência espacial), que geralmente não tem uma explicação física/científica, ou uma covariável referenciada no espaço (tendência de superfície).

A Figura 2.1 exemplifica o conceito de tendência espacial. Os dados[3] se referem à média mensal de precipitação em 143 estações no estado do Paraná durante os meses de maio e junho em diferentes anos. Observamos que a precipitação média varia de acordo com a localização e os gráficos sugerem uma tendência linear e/ou quadrática nas coordenadas.

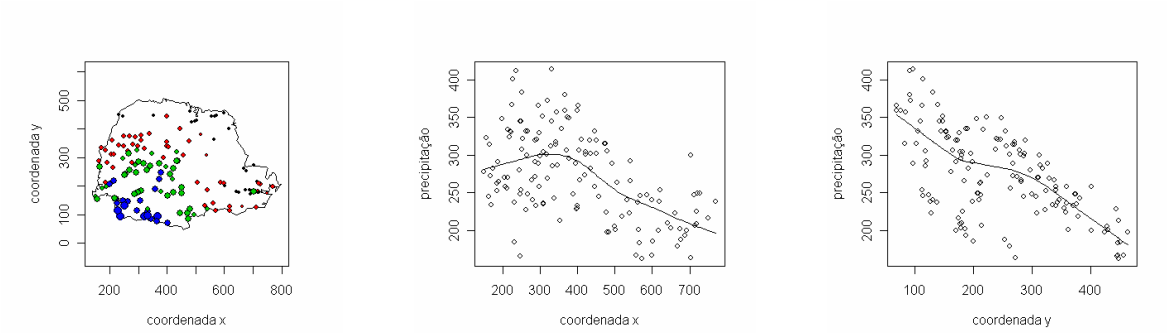


Figura 2.1. Gráfico de pontos da precipitação (esquerda), e gráficos de dispersão das coordenadas x (centro) e y (direita) versus a precipitação.

A variação da estrutura de covariância também pode originar um processo não estacionário. O leitor interessado em maiores informações deve verificar Schmidt e Sansó (2006). Estes autores fazem uma revisão das propostas de modelos com estruturas de covariâncias flexíveis, sob o enfoque bayesiano.

[3] Estes dados estão disponíveis no pacote geoR (2001) do *software* R (2006).

2.4 Variogramas

O variograma ou semi-variograma é uma ferramenta base na geoestatística que descreve a relação entre distância e dependência espacial. Esta função visa analisar a estrutura de variação das variáveis aleatórias no espaço.

O semi-variograma e a função de covariância, sob a hipótese de estacionariedade de segunda ordem do processo, são ferramentas equivalentes para descrever a dependência espacial entre duas variáveis aleatórias $Z(s+h)$ e $Z(s)$, separadas por uma distância h, como é demonstrado em (2.6).

$$\begin{cases} \gamma(h) = \frac{1}{2} Var[Z(s+h) - Z(s)] \\ \gamma(h) = \frac{1}{2}\{Var[Z(s+h)] + Var[Z(s)] - 2Cov[Z(s+h), Z(s)]\} \\ \gamma(h) = Cov(0) - Cov(h) \end{cases} \tag{2.6}$$

Observamos que o covariograma, quando plotamos a função de covariância versus a distância, é obtido segundo Armstrong (1998) virando o variograma para baixo, como ilustra a Figura 2.2.

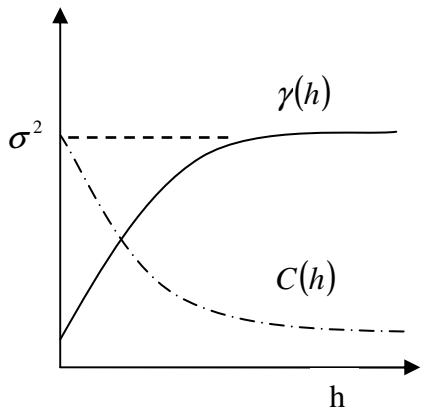


Figura 2.2. Covariograma e Variograma.

O correlograma é uma outra ferramenta que também pode ser utilizada para caracterizar a autocorrelação entre as variáveis, e é obtida como (JOURNEL; HUIJBREGTS, 1978):

$$\rho(h)=\frac{Cov(h)}{Cov(0)}=\frac{Cov(0)-\gamma(h)}{Cov(0)}=1-\frac{\gamma(h)}{Cov(0)}, \quad \text{e} \quad Cov(0)>0 \tag{2.7}$$

A justificativa do emprego extensivo do variograma na geoestatística em relação ao covariograma e ao correlograma se deve ao fato de nem sempre estarmos trabalhando com variância e covariância finitas *a priori* (JOURNEL; HUIJBREGTS, 1978), e a construção do variograma exige apenas a hipótese de estacionariedade intrínseca enquanto que o covariograma e o correlograma requerem uma suposição mais limitada, a de estacionariedade de segunda ordem.

As funções *madograma* e *rodograma*, propostas por Journel (1988) são ferramentas alternativas e pouco utilizadas que também descrevem a variabilidade espacial do processo. Mingoti (1996) mostra, usando três exemplos de aplicação, que as predições obtidas pela técnica de krigagem ordinária (interpolação espacial) quando empregamos essas duas medidas são melhores que aquelas obtidas com o variograma. Considere um processo estocástico intrinsecamente estacionário e isotrópico e seja:

$$E\left\{\left|Z(s_l)-Z(s_k)\right|^{w}\right\}=2\gamma_w\left(\left\|s_l-s_k\right\|\right), \quad \forall \quad s_l \neq s_k \in D \tag{2.8}$$

Para $w=1$ tem-se o m*adograma* e para $w=1/2$, o *rodograma* (MINGOTI, 1996). Se $w=2$ esta equação se reduz à equação (2.5), na qual obtemos o variograma.

Na análise espacial o variograma é estimado em duas etapas. Na primeira fase usamos os dados amostrais para calcular o variograma experimental (ou amostral), e na última ajustamos um modelo teórico ao variograma experimental e estimamos os seus parâmetros. A qualidade do ajuste do modelo pode ser avaliada pela validação cruzada.

A Figura 2.3 ilustra um variograma experimental típico e os seus parâmetros.

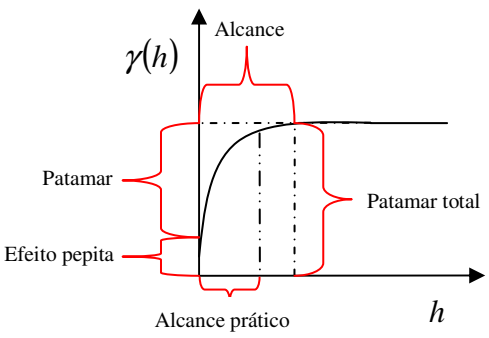


Figura 2.3. Representação esquemática de um variograma típico.

O variograma é uma medida de dissimilaridade, isto é, seu valor aumenta quando diminui a associação entre as variáveis. É esperado que observações mais próximas geograficamente tenham um comportamento mais semelhante entre si do que aquelas separadas por distâncias maiores.

Observamos pela Figura 2.3 que a função $\gamma(h)$ cresce até se estabilizar em um valor o qual denominamos de *patamar total* indicando que a partir deste ponto não existe mais dependência espacial entre as observações. A abscissa correspondente ao *patamar total* é chamada de *alcance*. Em alguns modelos teóricos de variograma o *patamar total* é atingido assintoticamente e neste caso define-se o *alcance prático* que segundo Schabenberger e Gotway (2005) é a distância (h) na qual $\gamma(h) = 0{,}95 \times \sigma^2$ (no caso de não ter efeito pepita), ou seja, é a distância na qual a correlação entre as observações é pequena e igual a 0,05.

Teoricamente o variograma deveria iniciar no valor zero ($\gamma(0) = 0$), pois esperamos que os valores observados nas mesmas localizações sejam iguais. Na prática, porém, isto nem sempre ocorre, pois à medida que *h* tende para zero, $\gamma(h)$ se aproxima de um valor positivo chamado *efeito pepita*. Este valor mostra a descontinuidade da função na origem, que é proveniente do erro de micro escala (ou variabilidade de pequena escala) não captada pela amostragem e/ou dos erros de medição. Conforme Armstrong (1998), o variograma pode apresentar outros três tipos de comportamento na origem que são: quadrático, linear e efeito pepita puro. O *patamar* é a diferença entre o *patamar total* e o *efeito pepita*.

2.4.1 Variogramas Experimentais

Existem diversos métodos para o cálculo do variograma experimental $(2\hat{\gamma}(h))$. O estimador clássico de Matheron baseado no método dos momentos (CRESSIE, 1993), é dado por:

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} \left[Z(s_i) - Z(s_j) \right]^2, h \in \Re^d \tag{2.9}$$

onde $N(h) = \left\{ (s_i, s_j), \| s_i - s_j \| = h; i, j = 1,2,...,n, i \neq j \right\}$ e $|N(h)|$ é o número de pares distintos em $N(h)$.

Quando substituímos o quadrado da expressão (2.9) por 1 ou por ½, obtemos respectivamente, o madograma e o rodograma amostrais (MINGOTI, 1996). Como ilustração,

apresentamos a seguir um exemplo de cálculo do variograma amostral pelo método dos momentos.

Exemplo 1: Considere um espaço unidimensional ($D = \Re^1$) onde as observações estão separadas por uma distância de dez unidades e estão dispostas da seguinte forma:

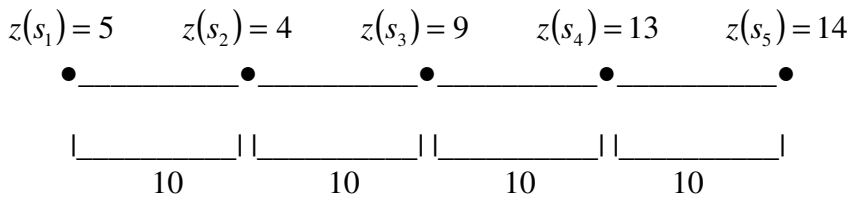


Figura 2.4. Dados no espaço unidimensional.

Temos que:

$$\begin{cases} 2\hat{\gamma}(10) = (5-4)^2 + (4-9)^2 + (9-13)^2 + (13-14)^2 / 4 = 10{,}75 \Rightarrow \hat{\gamma}(10) = 5{,}375. \\ 2\hat{\gamma}(20) = (5-9)^2 + (4-13)^2 + (9-14)^2 / 3 = 40{,}67 \Rightarrow \hat{\gamma}(20) = 20{,}335. \\ 2\hat{\gamma}(30) = (5-13)^2 + (4-14)^2 / 2 = 82 \Rightarrow \hat{\gamma}(30) = 41. \\ 2\hat{\gamma}(40) = (5-14)^2 = 81 \Rightarrow \hat{\gamma}(40) = 40{,}5. \end{cases}$$

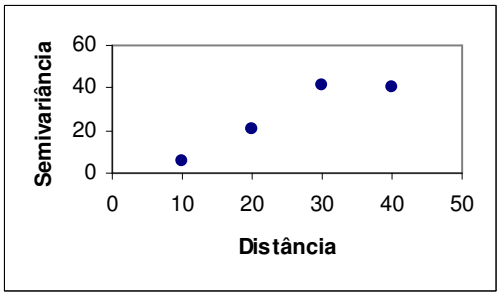


Gráfico 1. Semivariograma experimental.

Se os dados não estão distribuídos de forma regular devemos construir classes de distâncias. A regra de Journel e Huijbregts (1978) sugere que *h* seja menor que $L/2$, onde L é a maior distância entre as observações, e que o número de pares para cada h deva ser maior que 30. Rosa (2003) mostra que a escolha de valores de *h* próximos de 40% da distância máxima entre as observações é uma alternativa bem razoável.

No espaço bidimensional ainda temos que avaliar a tolerância angular. Armstrong (1998) diz que devemos calcular o variograma em pelo menos quatro direções distintas para validar ou não a suposição de isotropia.

Na literatura existem outros estimadores de variograma como o de Cressie e Hawkins (1980) e o de Genton (1998). Estes estimadores foram propostos com o objetivo de serem mais robustos em relação à presença de observações discrepantes (ou *outliers*) e a falta de normalidade de dados. Alguns trabalhos publicados comparam esses estimadores, entre eles o de Mingoti e Rosa (2008) que verificam que os estimadores de variograma da classe robusta fornecem bons resultados nas situações em que há *outliers*, sendo o estimador de Genton (1998) o que apresenta os melhores resultados. Na ausência de observações discrepantes, os autores observam que os estimadores da classe não robusta como os de Matheron definido em (2.9) e o das Diferenças de Hanslett (1997) são preferíveis.

2.4.2 Variogramas Teóricos

A segunda etapa de estimação do variograma consiste no ajuste de um modelo teórico ao variograma experimental (ou amostral). Os variogramas teóricos devem atender as seguintes propriedades (ARMSTRONG, 1998; SCHABENBERGER; GOTWAY, 2005; ROSA, 2003):

i. $2\gamma(0) = 0$;

ii. $\lim_{h\to\infty} 2\gamma(h) = 2\sigma^2$, onde $\sigma^2 = Var(Z(s_i)), \forall\ s_i \in D$;

iii. $2\gamma(h) = 2\gamma(-h)$;

iv. $2\gamma(h) \geq 0$ (é uma função não negativa);

v. $2\gamma(h)$ é um variograma válido se satisfazer à condição de função condicionalmente definida negativa, isto é:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j 2\gamma(h) \leq 0$$

para qualquer número finito de localizações $\{s_i, i = 1,2,...,k\}$ e quaisquer números reais $\{a_i, i = 1,2,...,k\}$, satisfazendo $\sum_{i=1}^{k} a_i = 0$.

A função de covariância espacial de um processo estocástico estacionário de segunda ordem também deve satisfazer a algumas propriedades descritas a seguir (ARMSTRONG, 1998; SCHABENBERGER; GOTWAY, 2005):

i. $C(0) = \sigma^2 \geq 0$;

ii. $C(h) = C(-h)$ (ou propriedade de simetria);

iii. $|C(h)| \leq C(0)$;

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

v. $C(h)$ é uma função de covariância válida se satisfazer a condição de positiva definida dada por:

$\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j C(s_i - s_j) \geq 0$ para quaisquer localizações s_i e s_j, e pesos $a_i, a_j; i, j = 1, 2, ..., k$.

vi. Se $C_j(h), j = 1, 2, ..., k$ é uma função de covariância válida, então a combinação linear de funções de covariância válidas também é uma função de covariância válida. Então, $\sum_{j=1}^{k} b_j C_j(h)$ é uma função de covariância válida se $b_j \geq 0, \forall j$.

vii. O produtório de funções de covariância válidas, também é uma função de covariância válida, i.e., se $C_j(h), j = 1, 2, ..., k$ é uma função de covariância válida, então $\prod_{j=1}^{k} C_j(h)$ também é uma função de covariância válida.

viii. Uma função de covariância que é válida em um espaço d - dimensional, é válida também em um espaço r - dimensional, com r < d.

Os principais modelos de semivariogramas teóricos para processos estacionários e isotrópicos são apresentados a seguir (ARMSTRONG, 1998; CRESSIE, 1993; SCHMIDT; SANSÓ, 2006):

- **Modelo Esférico**

$$\gamma(h) = \begin{cases} 0 & , h \leq 0 \\ \tau^2 + \psi^2 \left[\frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3 \right] & , 0 < h < a \\ \tau^2 + \psi^2 & , h \geq a \end{cases} \qquad (2.10)$$

O modelo esférico é valido em $\Re^1, \Re^2$ e $\Re^3$ e a abscissa correspondente ao patamar é dada por $\frac{2a}{3}$, onde *a* é o alcance (ou efeito escala). Temos que $\lim_{h \to 0^+} \gamma(h) = \tau^2$ é o efeito pepita e $\lim_{h \to \infty} \gamma(h) = \tau^2 + \psi^2 = \sigma^2$ é o patamar total. Este modelo tem um comportamento linear na origem.

- **Modelo Exponencial**

$$\gamma(h)=\begin{cases}\tau^2+\psi^2\left[1-\exp\left(\dfrac{-h}{a}\right)\right] & ,\ h>0\\ 0 & ,\ h\le 0\end{cases} \tag{2.11}$$

O modelo exponencial atinge o patamar assintoticamente e a distância correspondente é igual a *a* (alcance) ou igual a $1/3$ do alcance prático (a'). Quando $h=3a$ obtemos o alcance prático de 95%, pois $\gamma(3a)=\tau^2+\psi^2\left(1-\exp\left(\dfrac{-3a}{a}\right)\right)=\tau^2+\psi^2(0{,}95)$. O comportamento deste modelo na origem é linear.

- **Modelo Gaussiano**

$$\gamma(h)=\begin{cases}\tau^2+\psi^2\left\{1-\exp\left[-\left(\dfrac{h}{a}\right)^2\right]\right\} & ,\ h>0\\ 0 & ,\ h\le 0\end{cases} \tag{2.12}$$

O patamar para o modelo Gaussiano também é atingido assintoticamente e o alcance prático é dado por $a'=1{,}73a$. O comportamento deste modelo na origem é quadrático.

- **Mátern**

$$\gamma(h)=\begin{cases}\tau^2+\psi^2\left[1-\left(\dfrac{2\sqrt{k}h}{a}\right)^k\dfrac{1}{2^{k-1}\Gamma(k)}\mathrm{K}_k\left(\dfrac{2\sqrt{k}h}{a}\right)\right] & ,\ h>0\\ 0 & ,\ h\le 0\end{cases} \tag{2.13}$$

onde K_k denota a função de Bessel de ordem k.

- **Modelo Potência (função Poder)**

$$\gamma(h)=\begin{cases}\tau^2+\psi^2h^\alpha & ,\ h>0 \text{ e } 0<\alpha\le 2\\ 0 & ,\ h\le 0\end{cases} \tag{2.14}$$

O modelo linear é um caso particular da função potência quando $\alpha=1$, ou seja, $\gamma(h)=\tau^2+\psi^2h$ e é válido em $\Re^d, d\ge 1$. O comportamento na origem depende do valor de α.

- **Modelo Senoidal**

$$\gamma(h)=\begin{cases}\tau^2+\psi^2\left[1-a\dfrac{sen\left(\dfrac{h}{a}\right)}{h}\right] & ,\ h>0\\ 0 & ,\ h\le 0\end{cases} \tag{2.15}$$

O semivariograma senoidal apresenta correlação negativa originada da periodicidade do processo e é válido em $\Re^1, \Re^2$ e $\Re^3$ (CRESSIE, 1993). O patamar é atingido assintoticamente e o alcance prático é igual a três vezes o alcance, ou seja, $a' = 3a$. Apresenta um comportamento quadrático na origem.

- **Modelo Cauchy**

$$\gamma(h)=\begin{cases}\tau^2+\psi^2\left[1-\left(1+\left(\dfrac{h}{a}\right)^2\right)^{-k}\right] & ,\ h>0\\ 0 & ,\ h\le 0\end{cases} \tag{2.16}$$

- **Modelo Circular**

$$\gamma(h)=\begin{cases}\tau^2+\psi^2\left(\Gamma(h)\right) & ,h>0\\ 0 & ,h\le 0\end{cases}$$
$$\text{onde } \Gamma(h)=\frac{2\left\{\left(\theta\sqrt{1-\theta^2}\right)+sen^{-1}\sqrt{\theta}\right\}}{\pi} \text{ e } \theta=\min\left(\frac{h}{a},1\right) \tag{2.17}$$

- **Modelo Cúbico**

$$\gamma(h)=\begin{cases}0 & ,h\le 0\\ \tau^2+\psi^2\left[7\left(\dfrac{h}{a}\right)^2-8{,}75\left(\dfrac{h}{a}\right)^3+3{,}5\left(\dfrac{h}{a}\right)^5-0{,}75\left(\dfrac{h}{a}\right)^7\right] & ,0<h\le a\\ \tau^2+\sigma^2 & ,h>a\end{cases} \tag{2.18}$$

Este modelo apresenta um comportamento quadrático na origem.

- **Modelo Gencauchy (Cauchy generalizada)**

$$\gamma(h)=\begin{cases}\tau^2+\psi^2\left\{1-\left[1+\left(\frac{h}{a}\right)^{k_2}\right]^{\frac{-k_1}{k_2}}\right\} & ,h>0,k_1>0,0<k_2\le 2\\ 0 & ,h\le 0\end{cases} \tag{2.19}$$

- **Modelo *"Stable"* (Estável)**

$$\gamma(h)=\begin{cases}\tau^2+\psi^2\left\{1-\exp\left[\left(\frac{-h}{a}\right)^{k}\right]\right\} & ,h>0,0<k\le 2\\ 0 & ,h\le 0\end{cases} \tag{2.20}$$

- **Modelo Efeito Pepita Puro**

$$\gamma(h)=\begin{cases}\tau^2+\psi^2 k & ,h>0\\ 0 & ,h\le 0\end{cases} \tag{2.21}$$

onde *k* é uma constante qualquer. Este modelo representa um fenômeno completamente aleatório no qual não existe correlação espacial.

Como ilustração alguns gráficos de semivariogramas são apresentados na Figura 2.5.

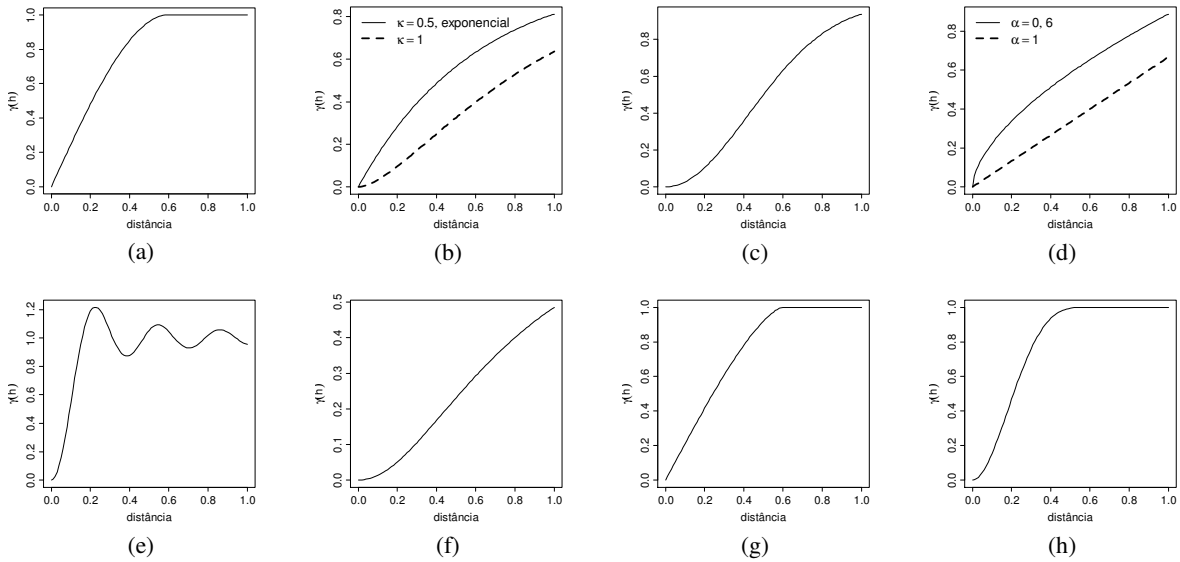


Figura 2.5. (a) Semivariograma esférico. (b) Semivariograma mátern. (c) Semivariograma gaussiano. (d) Semivariograma poder. (e) Semivariograma senoidal. (f) Semivariograma cauchy. (g) Semivariogram circular. (h) Semivariograma cúbico.

Outros modelos de semivariogramas teóricos podem ser encontrados em Chilés e Delfiner (1999), Journel e Huijbregts (1978) e Schabenberger e Gotway (2005).

A estimação dos parâmetros do variograma teórico pode ser feita utilizando o método de mínimos quadrados ou métodos baseados na função de verossimilhança. Diggle e Ribeiro Júnior (2007) tratam dos métodos de mínimos quadrados ordinários, mínimos quadrados ponderados e da máxima verossimilhança (verossimilhança restrita e verossimilhança perfilhada). Curriero e Lele (1999) estudam os métodos baseados na verossimilhança composta. A utilização de um determinado método depende de diversos fatores, como por exemplo, da distribuição da variável.

2.5 Métodos de Predição - Krigagem

A krigagem é um método de interpolação (ou predição) na geoestatística que pondera os vizinhos do ponto a ser estimado fornecendo estimativas pontuais não viciadas e de variância mínima. O termo "krigagem" foi dado por G. Matheron em 1963 em homenagem a D. G. Krige, um engenheiro de minas Sul Africano, que desenvolveu uma técnica de médias móveis para determinar a concentração de ouro em jazidas minerais (CRESSIE, 1993).

O interesse da geoestatística é predizer valores em locais não avaliados, como foi dito anteriormente. Para tanto, se queremos estimar um valor no ponto s_0 pode-se considerar um estimador que seja uma combinação linear dos dados na vizinhança de s_0 dado por (JOURNEL; HUIJBREGTS, 1978):

$$Z^*(s_0) = \sum_{i=1}^{k} \lambda_i Z(s_i) \tag{2.22}$$

onde $\lambda_i, i = 1, 2, ..., k$ são os pesos que devem ser escolhidos para obter um estimador que obedeça aos critérios:

1) Não viciado ou não tendencioso: $E\left[Z^*(s_0) - Z(s_0)\right] = 0$.
2) Variância mínima: $Var\left[Z^*(s_0) - Z(s_0)\right]$ é mínima, ou seja, é a menor possível.

Os tipos de krigagem mais usuais são a krigagem simples, a krigagem ordinária, a krigagem universal e a co-krigagem. Cressie (1993) também trata da krigagem bayesiana e Almeida e Ribeiro Júnior (1996) discutem a krigagem indicatriz. Descrevemos a seguir os procedimentos de krigagem ordinária e de krigagem simples.

2.5.1 Krigagem Ordinária

A krigagem ordinária é um método de estimação que é utilizado quando não conhecemos o valor da média espacial.

Considere um processo estocástico estacionário e isotrópico. Temos que $E\{Z(s)\} = m, \forall s \in D$ e isso implica que $E\{Z^*(s_0)\} = m$, onde $Z^*(s_0)$ é definido como em (2.22). A média do erro de estimação é obtida como:

$$E\left[Z^*(s_0) - Z(s_0)\right] = E\left[\sum_{i=1}^{k} \lambda_i Z(s_i) - Z(s_0)\right] = \sum_{i=1}^{k} \lambda_i m - m = m\left(\sum_{i=1}^{k} \lambda_i - 1\right) \tag{2.23}$$

Logo, para que o estimador seja não tendencioso temos que $m = 0$ ou $\sum_{i=1}^{k} \lambda_i = 1$. No entanto, como não conhecemos a média devemos supor que a soma dos pesos seja igual a um. Desta forma, obtemos um estimador não viciado.

A variância de estimação é dada por:

$$\begin{cases} Var\left[Z^*(s_0) - Z(s_0)\right] & = E\left\{\left[Z^*(s_0) - Z(s_0)\right]^2\right\} - \left\{E\left[Z^*(s_0) - Z(s_0)\right]\right\}^2 \\ & = E\left\{\left[Z^*(s_0) - Z(s_0)\right]^2\right\} \end{cases} \tag{2.24}$$

pois $E\left[Z^*(s_0) - Z(s_0)\right] = 0$, pela propriedade de não tendenciosidade.

A minimização da variância sujeita à restrição $\sum_{i=1}^{k} \lambda_i = 1$ e igualada a zero resulta no seguinte sistema de equações de krigagem dado por (MINGOTI et al., 2006):

$$\lambda_0 = \Gamma_0^{-1} \gamma_0 \tag{2.25}$$

sendo $\lambda_0 \equiv (\lambda_1, \lambda_2, ..., \lambda_k, \theta)'$, θ é o multiplicador de Lagrange, $\gamma_0 \equiv [\gamma(s_0 - s_1), ..., \gamma(s_0 - s_k), 1]$ e Γ_0^{-1} é a matriz inversa de Γ_0 dada por:

$$\Gamma_0 = \begin{cases} \gamma(s_i - s_j) & i = 1, ..., k; j = 1, ..., k \\ 1 & i = k+1; j = 1, ..., k \\ 0 & i = k+1; j = k+1 \\ 1 & i = 1, ..., k; j = k+1 \end{cases} \tag{2.26}$$

Na forma matricial temos:

$$\begin{bmatrix} \gamma(s_1-s_1) & \gamma(s_1-s_2) & \cdots & \gamma(s_1-s_k) & 1 \\ \gamma(s_2-s_1) & \gamma(s_2-s_2) & \cdots & \gamma(s_2-s_k) & 1 \\ \vdots & \vdots & \vdots & \vdots & 1 \\ \gamma(s_k-s_1) & \gamma(s_k-s_2) & \cdots & \gamma(s_k-s_k) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \hat{\lambda}_1 \\ \hat{\lambda}_2 \\ \vdots \\ \hat{\lambda}_k \\ \theta \end{bmatrix} = \begin{bmatrix} \gamma(s_0-s_1) \\ \gamma(s_0-s_2) \\ \vdots \\ \gamma(s_0-s_k) \\ 1 \end{bmatrix} \tag{2.27}$$

A solução do sistema de equações é dada por (CRESSIE, 1993, p. 122):

$$\begin{cases} \lambda' = \left(\gamma + 1 \dfrac{\left(1 - 1'\Gamma^{-1}\gamma\right)}{1'\Gamma^{-1}1} \right)' \Gamma^{-1} \\ \theta = -\left(1 - 1'\Gamma^{-1}\gamma\right) \Big/ \left(1'\Gamma^{-1}1\right) \end{cases} \tag{2.28}$$

onde $\gamma \equiv (\gamma(s_0 - s_1), ..., \gamma(s_0 - s_k))'$ e Γ é a matriz $k \times k$ cujo elemento (i, j) é $\gamma(s_i - s_j)$.

A estimativa de $Z(\cdot)$ na localização não amostrada é obtida como:

$$z^*(s_0) = \sum_{i=1}^{k} \hat{\lambda}_i z(s_i) \tag{2.29}$$

Se estivermos interessados em recuperar uma superfície, estimamos uma malha de pontos que devem cobrir toda a região e pertencer ao domínio D. Dessa maneira, podemos avaliar o comportamento espacial da característica investigada.

A seguir apresentamos um exemplo deste procedimento de estimação. Considere uma realização do processo estocástico $Z = \{Z(s), s \in D \subset \Re^2\}$, no qual os valores da realização deste processo, supondo que estamos trabalhando com quatro variáveis aleatórias, e as respectivas coordenadas são dados pela Tabela 2.1.

Tabela 2.1 - Valores observados e localizações.

Observação	x	y	Valor
$z(s_1)$	0,58	0,82	0,72
$z(s_2)$	0,22	0,42	0,66
$z(s_3)$	0,33	0,96	0,44
$z(s_4)$	0,71	0,98	0,67

Suponha que queiramos predizer o valor em um local não amostrado com coordenadas $s_0 = (x, y)$ iguais a (0,44; 0,79). A distribuição espacial dos dados observados (círculos) e do ponto a ser predito (triângulo) é ilustrada pela Figura 2.6.

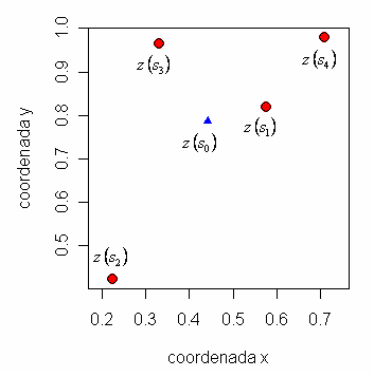


Figura 2.6. Distribuição espacial das amostras.

A matriz de distâncias *H* entre as observações $z(s_1), z(s_2), z(s_3)$ e $z(s_4)$ é dada por (2.30):

$$H = \begin{bmatrix} 0 & 0{,}538 & 0{,}287 & 0{,}206 \\ 0{,}538 & 0 & 0{,}551 & 0{,}744 \\ 0{,}287 & 0{,}551 & 0 & 0{,}381 \\ 0{,}206 & 0{,}744 & 0{,}381 & 0 \end{bmatrix} \tag{2.30}$$

As distâncias entre as observações $z(s_1), z(s_2), z(s_3), z(s_4)$ e o ponto de predição (s_0) são dispostas no vetor *V*:

$$V = \begin{bmatrix} 0{,}143 & 0{,}430 & 0{,}202 & 0{,}330 \end{bmatrix} \tag{2.31}$$

Apenas como ilustração, suponha um modelo teórico de semivariograma linear igual a $\gamma(h) = \tau^2 + \psi^2 h = 0{,}25h$, i.e., não temos o efeito pepita e a variância é igual a 0,25. Substituindo os valores da matriz *H* e do vetor *V* na equação do modelo de variograma obtemos o sistema de equações de krigagem dado em (2.27) que na forma matricial pode ser escrito como em (2.32).

$$\begin{bmatrix} 0 & 0{,}135 & 0{,}072 & 0{,}052 & 1 \\ 0{,}135 & 0 & 0{,}138 & 0{,}186 & 1 \\ 0{,}072 & 0{,}138 & 0 & 0{,}095 & 1 \\ 0{,}052 & 0{,}186 & 0{,}095 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \hat{\lambda}_1 \\ \hat{\lambda}_2 \\ \hat{\lambda}_3 \\ \hat{\lambda}_4 \\ \theta \end{bmatrix} = \begin{bmatrix} 0{,}036 \\ 0{,}108 \\ 0{,}051 \\ 0{,}083 \\ 1 \end{bmatrix} \tag{2.32}$$

onde $\gamma(s_1 - s_2) = 0{,}25 \times h = 0{,}25 \times 0{,}538 = 0{,}135$, $\gamma(s_0 - s_1) = 0{,}25 \times h = 0{,}25 \times 0{,}143 = 0{,}036$ e assim por diante.

Os parâmetros estimados pela solução deste sistema são dados por (2.33):

$$\begin{bmatrix} \hat{\lambda}_1 \\ \hat{\lambda}_2 \\ \hat{\lambda}_3 \\ \hat{\lambda}_4 \\ \theta \end{bmatrix} = \begin{bmatrix} 0{,}56 \\ 0{,}15 \\ 0{,}34 \\ -0{,}04 \\ -0{,}01 \end{bmatrix} \tag{2.33}$$

Logo, o valor predito usando os valores de $Z(\cdot)$ dados na Tabela 2.1 será:

$$z^*(s_0) = \sum_{i=1}^{4} \hat{\lambda}_i z(s_i) = (0{,}56 \times 0{,}72 + \ldots - 0{,}04 \times 0{,}67) = 0{,}625 \tag{2.34}$$

2.5.2 Krigagem Simples

A krigagem simples supõe que o valor da média de $Z(\cdot)$ é conhecido. Considere um processo estocástico intrinsecamente estacionário com média conhecida e igual à zero. Então, tem-se que $E[Z(s)] = E[Z^*(s_0)] = 0, \forall s \text{ e } s_0 \in D$. Neste caso, não precisamos supor que a soma dos pesos seja igual a 1 (ver equação 2.23) para obter um estimador não viciado, pois se a média é zero isto implica que a esperança do erro de estimação também é zero, como demonstrada a seguir:

$$E[Z^*(s_0) - Z(s_0)] = E\left[\sum_{i=1}^{k} \lambda_i Z(s_i) - Z(s_0)\right] = \left(\sum_{i=1}^{k} \lambda_i \times 0\right) - 0 = 0 \tag{2.35}$$

A variância de estimação é dada como em (2.24). Armstrong (1998) mostra que a minimização da variância do erro de predição resulta em um sistema do tipo:

$$\sum_{i=1}^{k} \hat{\lambda}_i \gamma(s_i - s_j) = \gamma(s_0 - s_j), \quad j = 1, 2, ..., k \tag{2.36}$$

Este sistema não precisa do multiplicador de Lagrange, pois não é preciso que a soma dos pesos seja igual a 1. A solução do sistema resulta na seguinte equação de predição:

$$z^*(s_0) = \sum_{i=1}^{k} \hat{\lambda}_i z(s_i) \tag{2.37}$$

Se a média é diferente de zero, obtemos (ARMSTRONG, 1998):

$$z^*(s_0) = \sum_{i=1}^{k} \hat{\lambda}_i z(s_i) + m\left[1 - \sum_{i=1}^{k} \hat{\lambda}_i\right] \tag{2.38}$$

onde *m* é a média do processo.

2.6 Validação Cruzada

A validação cruzada é um procedimento para avaliar se o modelo de predição ajustado aos dados é adequado. A validação pode ser realizada de duas maneiras. Uma forma é retirar ponto a ponto do banco de dados e predizê-los usando o modelo espacial ajustado com os pontos restantes e em seguida analisar os resíduos. Na segunda forma selecionamos amostras de pontos de *D* e predizemos estas observações usando o modelo geoestatístico ajustado e depois fazemos à avaliação dos resíduos. Neste caso a seleção pode ser realizada várias vezes para retirar o efeito de amostragem e é útil quando temos um banco de dados muito grande.

Se o modelo variográfico ajustado aos dados for adequado espera-se que os resíduos tenham média próxima de zero e uma distribuição aproximadamente normal. Na validação cruzada é comum utilizar o resíduo padronizado que é dado por:

$$\hat{r}_i^P = \frac{\left(z(s_i) - z^*(s_i)\right)}{\hat{\sigma}_Z} \tag{2.39}$$

onde $\hat{\sigma}_Z$ é o desvio–padrão do erro de predição. Espera-se que os resíduos padronizados tenham média igual a zero e a variância igual a 1.

Capítulo 3

Modelos Espaço–Temporais: Geoestatística e Séries Temporais

Neste capítulo definimos o conceito de função aleatória espaço–temporal e discutimos os tipos de representações para caracterizar processos desta natureza e em especial, a função de covariância. Apresentamos a família de funções de covariância proposta por Gneiting (2002), o modelo geoestatístico de Høst et al. (1997) e o modelo de séries temporais de Niu et al. (2003). Também sugerimos algumas modificações nos modelos de Høst et al. (1995) e de Niu et al. (2003) e, além disso, propomos uma estratégia de modelagem que combina esses modelos modificados em duas etapas distintas de análise.

3.1 Conceitos Iniciais

Os procedimentos de estatística, freqüentemente, não são suficientes para descrever os processos espaço–temporais, pois não consideram a interação entre o espaço e o tempo, ou seja, estes procedimentos não conseguem captar a variabilidade nas dimensões espaço e tempo conjuntamente. As ferramentas para a análise de dados puramente temporais ou dados puramente espaciais são bem conhecidas, porém os métodos que consideram a dependência espaço–temporal são recentes e a modelagem deste tipo de processo é uma das áreas, atualmente, segundo Huang et al. (2007) de maior crescimento com diversas aplicações em ciências ambientais, ciências geofísicas, biologia, epidemiologia e outras. A dificuldade em modelar processos correlacionados no espaço e no tempo é motivada, em parte, pela necessidade de modelos de covariância não-separáveis, pois estes modelos requerem que as funções de covariância sejam válidas atendendo a condição de serem positivas definidas (MA, 2003).

Schabenberger e Gotway (2005) enumeram três procedimentos para a análise de processos espaço–temporais. Dois destes procedimentos, denominados métodos condicionais, analisam os dados do processo utilizando técnicas usuais de séries temporais (ou de estatística espacial) em cada localização (ou para cada tempo separadamente), ou seja, desconsidera-se a dependência espacial (ou temporal). Uma das desvantagens deste procedimento é que a análise separada do tempo (ou espaço) não permite predições em tempos futuros (ou novas

localizações). Os métodos condicionais são técnicas que podem ser utilizadas em uma análise exploratória de dados (EDA) espaço–temporais.

O terceiro procedimento, citado por Schabenberger e Gotway (2005), considera o tempo como uma dimensão espacial extra, ou seja, a análise é feita utilizando métodos para campos aleatórios definidos em $\Re^{d+1}$, onde d é a dimensão espacial. Segundo Kyriakidis e Journel (1999), em uma revisão abrangente sobre processos espaço–temporais, a distância e o tempo não são grandezas comparáveis. No tempo existe uma idéia clara de ordenação temporal (passado, presente e futuro) que não pode ser definida no espaço. A isotropia, conceito bem definido no espaço, não tem significado na dimensão tempo. A variação no processo espacial é diferente do processo temporal, pois a métrica nas duas dimensões não é a mesma, ou seja, as distâncias calculadas entre pontos no espaço e entre pontos no tempo não podem ser comparadas.

De acordo com Gneiting (2002, p. 591),

> Da perspectiva matemática, não existe distinção entre o domínio espaço–temporal $\Re^d \times \Re$ e o domínio puramente espacial $\Re^{d+1}$. Em outras palavras, a classe de funções de covariância espaço–temporais em $\Re^d \times \Re$ coincide com a classe de funções de covariância espacial $\Re^{d+1}$. Então, a diferença física fundamental entre as dimensões espacial e temporal deve ser reconhecida através da nossa notação e através de construções específicas [...].

Então as funções de covariância devem refletir essa diferença (física) de grandezas nas matrizes de anisotropia (GNEITING; SCHLATHER, 2002; SCHABENBERGER; GOTWAY, 2005). A matriz de anisotropia de processos espaço-temporais definidos em $\Re^2 \times \Re$ é descrita na seção 3.1.3.

Na próxima seção é explicado o conceito de função aleatória espaço–temporal sendo as condições de estacionariedade e de validade de funções de covariância redefinidas para acomodar a dimensão temporal.

3.1.1 Função Aleatória Espaço–Temporal

Seja $\{Z(s,t)\}$ a variável aleatória espaço–temporal medida na localização $s \in D(t) \subset \Re^d$ e no tempo $t \in T \subset \Re^1$. Temos que $D(t)$ é o domínio espacial dependente do

tempo, T é o domínio finito do tempo e $z(s,t),(s,t)\in D(t)\times T$, é uma realização da variável espaço–temporal. Nesta dissertação vamos supor que a região de estudo D não varia no tempo, ou seja, $D(t)\equiv D$ e que estamos trabalhando com processos espaciais bidimensionais ($d=2$).

A variável aleatória indexada no tempo e no espaço é caracterizada por sua distribuição de probabilidade acumulada (ou função de distribuição) definida como (KYRIAKIDIS; JOURNEL, 1999):

$$F(s,t;z)=P\{Z(s,t)\le z\},\ \forall z,(s,t)\in D\times T \tag{3.1}$$

O conjunto destas variáveis aleatórias, usualmente dependentes, determina a função aleatória espaço–temporal (ou processo estocástico espaço–temporal) dada por:

$$Z=\{Z(s,t),(s,t)\in D\times T\} \tag{3.2}$$

Se a condição de regularidade $(\mathrm{var}(Z(s,t))<\infty,\ \forall(s,t)\in D\times T)$ é verificada, então os dois primeiros momentos do processo existem. A média e a função de covariância do processo espaço-temporal são obtidas como (MA, 2002, 2005):

$$\begin{aligned}&\mu(s,t)=E\{Z(s,t)\}\\&C(s_1,s_2;t_1,t_2)=E\{Z(s_1,t_1)-E[Z(s_1,t_1)]\}\times E\{Z(s_2,t_2)-E[Z(s_2,t_2)]\}\\&\qquad=\mathrm{cov}(Z(s_1,t_1),Z(s_2,t_2)),\quad (s_1,t_1),(s_2,t_2)\in D\times T\end{aligned} \tag{3.3}$$

O variograma é dado por (MA, 2003):

$$\gamma(s_1,s_2;t_1,t_2)=\frac{1}{2}\mathrm{var}(Z(s_1,t_1)-Z(s_2,t_2)),\ (s_1,t_1),(s_2,t_2)\in D\times T \tag{3.4}$$

O processo estocástico espaço–temporal Z definido em (3.2) é descrito pela lei espaço–temporal obtida como (KYRIAKIDIS; JOURNEL, 1999):

$$F\{s_1,t_1,...,s_k,t_n;z_{11},...,z_{kn}\}=P\{Z(s_1,t_1)\le z_{11},...,Z(s_k,t_n)\le z_{kn}\} \tag{3.5}$$

Isto significa que devemos conhecer a distribuição conjunta de $\{Z(s_1,t_1),Z(s_2,t_2),...,Z(s_k,t_n)\}$ para qualquer inteiro positivo *k*, *n* e qualquer configuração dos pares de coordenadas espaço–temporais $(s_1,t_1),(s_2,t_2),...,(s_k,t_n)$.

Segundo Kyriakidis e Journel (1999) "a inferência da lei espaço–temporal requer realizações repetidas das variáveis aleatórias $Z(s,t)$ em cada localização espaço-temporal

$(s,t) \in D \times T$, o que é raramente disponível na prática". Dessa forma os pares de observações separados pelo vetor de distância *h* e pelo vetor de tempo *u* serão considerados medidas repetidas do processo. Algumas suposições sobre a estabilidade do processo espaço–temporal devem então ser consideradas.

A estacionariedade espaço–temporal estrita diz que a função de distribuição *kn* variada é invariante a translação, isto é (KYRIAKIDIS; JOURNEL, 1999):

$$\begin{aligned} &F\{s_1,t_1,...,s_k,t_n;z_{11},...,z_{kn}\} = F\{s_1+h,t_1+u,...,s_k+h,t_n+u;z_{11},...,z_{kn}\}, \\ &\forall\, s_1,t_1,...,s_k,t_n \text{ e } (h,u) \in D \times T \end{aligned} \tag{3.6}$$

O campo aleatório é (fracamente) estacionário no espaço se $E\{Z(s_1,t)\} = E\{Z(s_2,t)\}$ e a $Cov(s_1,s_2;t_1,t_2)$ depende de s_1 e s_2 somente através do vetor de separação espacial $h = \| s_1 - s_2 \|$, $\forall\, t_1,t_2 \in T$. O campo aleatório é dito (fracamente) estacionário no tempo se $E\{Z(s,t_1)\} = E\{Z(s,t_2)\}$ e a função de covariância depende somente do *lag* temporal $u = |t_1 - t_2|$ (HUANG et al., 2007).

A hipótese de estacionariedade espaço–temporal de segunda ordem (ou fraca), que envolve apenas os dois primeiros momentos da função aleatória $Z(s,t)$, diz que a média é constante ao longo de $D \times T$ e que a função de covariância depende somente do vetor de separação espaço–temporal, $(h,u) = (s_1 - s_2, t_1 - t_2)$, e não da localização (s_1,s_2) ou dos tempos (t_1,t_2), então:

$$\begin{aligned} &E\{Z(s_1,t_1)\} = E\{Z(s_2,t_2)\} = m, \quad \forall (s_1,t_1) \text{ e } (s_2,t_2) \in D \times T \\ &Cov(Z(s_1,t_1),Z(s_2,t_2)) = Cov(s_1 - s_2, t_1 - t_2) = C(h,u), \quad \forall (s_1,t_1) \text{ e } (s_2,t_2) \in D \times T \end{aligned} \tag{3.7}$$

onde $h = \| s_1 - s_2 \|$ e $u = |t_1 - t_2|$, $(h,u) \in D \times T$.

A função de covariância $C(h,u)$ é particularmente informativa para campos aleatórios gaussianos, pois estes são completamente especificados pelo vetor de médias e pela função de covariância (GNEITING; SCHLATHER, 2002).

A estacionariedade de segunda ordem espaço–temporal implica na estacionariedade intrínseca espaço–tempo. Esta última é baseada na função de variograma e diz que (MA, 2003):

$$
\begin{aligned}
&E\{Z(s_1,t_1)-Z(s_2,t_2)\}=0, \quad \forall(s_{i1},t_i)\in D\times T, i=1,2\\
&\gamma(s_1,s_2;t_1,t_2)=\frac{E[Z(s_1,t_1)-Z(s_2,t_2)]^2}{2}\\
&\qquad\qquad =C(0,0)-C(h,u)=\gamma(h,u), \quad \forall(s_i,t_i),(h,u)\in D\times T, i=1,2
\end{aligned}
\tag{3.8}
$$

onde $C(0,0)$ é a variância do processo.

A função de covariância estacionária é isotrópica se esta é invariante em relação à rotação e a translação (PORCU et al., 2007):

$$C(h,u)=\overline{C}(\|h\|,|u|), \quad (h,u)\in \Re^d\times\Re \tag{3.9}$$

onde $\|\cdot\|$ denota a norma Euclidiana e $\overline{C}$ é uma função positiva definida.

A função de covariância do processo estacionário $C(h,u)$ deve satisfazer a condição de ser *positiva definida* para que seja considerada uma função de covariância válida. Porcu et al. (2007) denominam essa condição de *permissibilidade*. Isso implica que para qualquer conjunto de pontos $(s_1,t_1)...(s_k,t_n)\in \Re^d\times\Re$ e coeficientes $a_i,a_j\in\Re, i=1,2,...,k$ e $j=1,2,...,n$, a condição (3.10) deve ser satisfeita.

$$\sum_{i=1}^{k}\sum_{j=1}^{n}a_i a_j C(s_i-s_j,t_i-t_j)\geq 0 \tag{3.10}$$

3.1.2 Representação do Processo Estocástico Espaço–Temporal

Segundo Huang et al. (2007, p. 4578) "o campo aleatório espaço–temporal é geralmente descrito de três maneiras distintas: uma representação estocástica dinâmica, uma função de covariância espaço–temporal, e uma função de densidade espectral espaço–temporal".

A função de densidade espectral espaço–temporal é particularmente importante quando não existem expressões de forma fechada para a função de covariância. Este procedimento é baseado no teorema de Bochner[4] que diz que uma função de covariância válida tem uma representação espectral. Isto sugere que a função de covariância válida pode ser obtida a partir da inversão da transformada de Fourier de uma função de densidade espectral válida (SCHABENBERGER; GOTWAY, 2005). Cressie e Huang (1999) propõem

[4] Outras informações sobre o teorema de Bochner podem ser obtidas em Bochner (1955).

um procedimento através da inversão de Fourier na forma fechada para obter funções de covariância permissíveis (GNEITING; SCHLATHER, 2002). Stein (2005) considera uma classe paramétrica de densidades espectrais cujas funções de covariância são infinitamente diferenciáveis longe da origem permitindo diferentes graus de suavidade do processo no espaço e no tempo. Gneiting (2002) generaliza o procedimento de Cressie e Huang (1999) usando funções monótonas e funções onde a primeira derivada é completamente monótona. O modelo proposto por Gneiting (2002) é tratado com maiores detalhes na seção 3.2.

A representação estocástica dinâmica considera equações diferenciais estocásticas para descrever a função aleatória espaço–temporal. Segundo Huang et al. (2007) a equação diferencial ordinária ou parcial descreve modelos determinísticos e os modelos estocásticos ou dinâmicos são definidos pelas equações diferenciais estocásticas.

Nesta dissertação abordaremos somente as funções de covariância para descrever o processo estocástico espaço–temporal e estas serão discutidas com maiores detalhes na próxima seção.

3.1.3 Funções de Covariância

As funções de covariância espaço–temporais podem ser separáveis ou não-separáveis. A função de covariância separável é caracterizada pela fatoração da função de covariância em duas componentes, uma componente puramente espacial e a outra puramente temporal. Funções separáveis de covariância são simples de serem obtidas, visto que a combinação linear ou o produto de funções de covariância válidas implicam em uma função de covariância válida (ver seção 2.4.2, p. 23). A decomposição do modelo separável, no caso aditivo, pode ser escrita como:

$$Cov(Z(s_1,t_1),Z(s_2,t_2)) = Cov(Z(s_1,s_2)) + Cov(Z(t_1,t_2)) = C(h;\theta_s) + C(u;\theta_t) \tag{3.11}$$

onde $C(h;\theta_s)$ e $C(u;\theta_t)$ são respectivamente as funções de covariância puramente espacial e puramente temporal e θ_s e θ_t são os parâmetros no espaço e no tempo associados a cada uma destas funções. No caso multiplicativo a função é da forma:

$$Cov(Z(s_1,t_1),Z(s_2,t_2)) = Cov(Z(s_1,s_2)) \times Cov(Z(t_1,t_2)) = C(h;\theta_s)C(u;\theta_t) \tag{3.12}$$

A fatoração da função de covariância espaço–temporal é feita aplicando-se as operações de adição e de multiplicação. Segundo Kyriakidis e Journel (1999, p. 664) o

modelo separável aditivo “no jargão de geoestatística, corresponde ao modelo de anisotropia zonal”. Os modelos separáveis são computacionalmente preferíveis aos modelos não separáveis.

“Qualquer função de covariância espaço–temporal que não pode ser escrita na forma de (3.11 ou 3.12) é dita não-separável” (GNEITING; SCHLATHER, 2002, p. 2). Alguns modelos de covariância não-separável se reduzem aos modelos separáveis para valores específicos dos parâmetros.

A separabilidade é uma propriedade conveniente, pois facilita a obtenção de funções de covariância válidas. Uma das grandes desvantagens dos modelos separáveis é que estes não conseguem incorporar a interação espaço–tempo. O processo conjunto (espaço e tempo) é modelado como dois processos independentes, um processo espacial e o outro temporal. A aplicação de modelos separáveis, geralmente, não tem uma explicação física e segundo Gneiting e Schlather (2002) são ruins para caracterizar processos naturais.

Outra propriedade das funções de covariância é a de simetria completa definida como (GNEITING, 2002):

$$C(h,u) = C(-h,u) = C(h,-u) = C(-h,-u), \quad (h,u) \in \Re^d \times \Re \tag{3.13}$$

Essa propriedade não é satisfeita para processos atmosféricos, ambientais e geofísicos, pois estes são influenciados por correntes oceânicas ou pela prevalência de direção do vento ao longo do tempo (GNEITING, 2002).

Segundo Silva (2006) a matriz de anisotropia de processos espaço-temporais pode ser especificada como:

$$A_i = \begin{pmatrix} a_{ixx} & a_{ixT} \\ a_{iTx} & a_{iTT} \end{pmatrix} \tag{3.14}$$

onde os parâmetros de escala que determinam a suavidade do processo para o espaço e para o tempo são respectivamente, a_{ixx} e a_{iTT}. Os parâmetros de forma para a dimensão espacial e para a dimensão temporal correspondem a respectivamente, a_{ixT} e a_{iTx}. Para processos espaciais de dimensão maior que 1 ($D \in \Re^d, d > 1$), a_{ixx} é uma matriz de dimensão $d \times d$. Um processo é simétrico se os elementos fora da diagonal principal da matriz de anisotropia são iguais à zero. Quando consideramos funções de covariância separáveis, obtemos duas matrizes de anisotropia (A_i e A_j) uma referente à função de covariância puramente espacial e

a outra referente à covariância puramente temporal. Outras informações e exemplos sobre a matriz de anisotropia podem ser vistos em Silva (2006).

O *software* estatístico R (2006) possui um pacote chamado *RandomFields* (2006) desenvolvido pelo pesquisador Martin Schlather, que permite a análise e a simulação de processos espaço–temporais. Silva (2006) utiliza a função *GaussRF()* deste pacote para simular processos espaço–temporais com diferentes características quanto à separabilidade, a simetria e a isotropia espacial.

A seção a seguir descreve a família de funções de covariância proposta por Gneiting (2002). Na análise dos dados que será apresentada nos Capítulos 4, 5 e 6 ajustamos um modelo que é um caso particular desta família de funções de covariância espaço-temporal.

3.2 Funções de Covariância Espaço-Temporal Propostas por Gneiting (2002)

Cressie e Huang (1999) introduzem um procedimento, baseado na inversão de Fourier e que requer operações no domínio espectral, que permite obter uma classe de funções de covariância não-separável espaço–temporal para processos estacionários. Este procedimento produz expressões fechadas e funções de covariância válidas.

A proposta de Gneiting (2002) é uma generalização do procedimento de Cressie e Huang (1999), porém não depende da inversão de Fourier na forma fechada, ou seja, a construção da classe de funções de covariância não separável espaço–temporal estacionária é direta realizando-se no domínio espaço–temporal. As funções de covariância são formadas a partir de componentes elementares cuja validade é facilmente verificada. Seja $\varphi(t), t \geq 0$ uma função completamente monótona e $\psi(t)$ uma função positiva com derivada completamente monótona. De acordo com Gneiting e Schlather (2002): "$\varphi(t)$, $t \geq 0$ é chamada completamente monótona se possui derivadas $\varphi^{(n)}$ de todas as ordens e $(-1)^n \varphi^{(n)}(t) \geq 0$ para $t > 0$ e $n = 0,1,2...$". As Tabelas 3.1 e 3.2 mostram alguns exemplos de funções completamente monótonas e funções positivas com derivadas completamente monótonas. Estas tabelas são reproduções daquelas apresentadas no trabalho de Gneiting (2002).

Tabela 3.1 - Funções completamente monótonas*

Função	Parâmetros
$\varphi(t) = \exp\left(-ct^{\gamma}\right)$	$c > 0, 0 < \gamma \leq 1$
$\varphi(t) = \left(2^{v-1}\Gamma(v)\right)^{-1}\left(ct^{1/2}\right)^{v} \mathrm{K}_{v}\left(ct^{1/2}\right)$	$c > 0, v > 0$
$\varphi(t) = \left(1 + ct^{\gamma}\right)^{v}$	$c > 0, 0 < \gamma \leq 1, v > 0$
$\varphi(t) = 2^{v}\left(\exp\left(ct^{1/2}\right) + \exp\left(-ct^{1/2}\right)\right)^{v}$	$c > 0, v > 0$

onde K_{v} denota a função de Bessel modificada do segundo tipo de ordem v.

Tabela 3.2 - Funções positivas com derivadas completamente monótonas*

Função	Parâmetros
$\psi(t) = \left(at^{\alpha} + 1\right)^{\beta}$	$a > 0, 0 < \alpha \leq 1, 0 \leq \beta \leq 1$
$\psi(t) = \dfrac{\ln\left(at^{\alpha} + b\right)}{\ln(b)}$	$a > 0, b > 1, 0 < \alpha \leq 1$
$\psi(t) = \dfrac{\left(at^{\alpha} + b\right)}{\left(b\left(at^{\alpha} + 1\right)\right)}$	$a > 0, 0 < b \leq 1, 0 < \alpha \leq 1$

A função de covariância espaço–temporal válida em $\Re^{d} \times \Re$ é dada por (GNEITING, 2002):

$$C(h,u) = \frac{\sigma^{2}}{\psi\left(|u|^{2}\right)^{d/2}} \varphi\left(\frac{\|h\|^{2}}{\psi\left(|u|^{2}\right)}\right), \quad (h,u) \in \Re^{d} \times \Re \tag{3.15}$$

onde $\|h\|$ é a distância espacial, $|u|$ é a distância temporal, $\varphi(t), t \geq 0$ é uma função completamente monótona e $\psi(t)$ uma função positiva com derivada completamente monótona, como aquelas apresentadas nas Tabelas 3.1 e 3.2. As funções $\varphi(t)$ e $\psi(t)$ estão

* Reprodução de Gneiting (2002)

geralmente, relacionadas com as estruturas, respectivamente espacial e temporal; $\sigma^2 > 0$ é a variabilidade do processo e $d > 0$ é a dimensão espacial.

Se tomarmos $d = 2$, $\varphi(t) = \exp\left(-ct^{\gamma}\right)$, $c > 0$, $0 < \gamma \leq 1$ e $\psi(t) = \left(at^{\alpha} + 1\right)^{\beta}$, $a > 0$, $0 < \alpha \leq 1$, $0 \leq \beta \leq 1$ e substituirmos na expressão (3.15) obtemos:

$$\begin{aligned} C(h,u) &= \frac{\sigma^2}{\psi\left(|u|^2\right)}\varphi\left(\frac{\|h\|^2}{\psi\left(|u|^2\right)}\right) = \frac{\sigma^2}{\left(a|u|^{2\alpha}+1\right)^{\beta}}\varphi\left(\frac{\|h\|^2}{\left(a|u|^{2\alpha}+1\right)^{\beta}}\right) \\ &= \frac{\sigma^2}{\left(a|u|^{2\alpha}+1\right)^{\beta}}\exp\left(-c\frac{\|h\|^{2\gamma}}{\left(a|u|^{2\alpha}+1\right)^{\beta\gamma}}\right), \quad a,c,\sigma^2 > 0 \text{ e } \alpha,\gamma,\beta \in (0,1] \end{aligned} \tag{3.16}$$

sendo a e α os parâmetros de escala e de suavidade do processo no tempo, e c e γ os parâmetros de escala e de suavidade do processo no espaço (GNEITING, 2002). Se o parâmetro β for igual à zero, obtemos uma função de covariância puramente espacial, ou seja, esta não depende do vetor de separação temporal u como vemos em (3.17):

$$C(h,u;\beta = 0) = \sigma^2 \exp\left(-c\|h\|^{2\gamma}\right) \tag{3.17}$$

A multiplicação da função de covariância em (3.16) por uma função de covariância puramente temporal, $C(u) = \left(a|u|^{2\alpha}+1\right)^{-\delta}$, resulta em uma função de covariância espaço–temporal válida não-separável (SCHABENBERGER; GOTWAY, 2005):

$$C(u) \times C(h,u) = \frac{\sigma^2}{\left(a|u|^{2\alpha}+1\right)^{\delta+\beta}}\exp\left(-c\frac{\|h\|^{2\gamma}}{\left(a|u|^{2\alpha}+1\right)^{\beta\gamma}}\right) \tag{3.18}$$

Se $\beta = 0$ em (3.18) obtemos uma função de covariância espaço–temporal válida separável da forma $C(h,u) = \sigma^2 C(u)C(h)$ dada por:

$$C(h,u;\beta = 0) = \frac{\sigma^2}{\left(a|u|^{2\alpha}+1\right)^{\delta}}\exp\left(-c\|h\|^{2\gamma}\right) \tag{3.19}$$

Na expressão (3.19) temos que $C(h) = \exp\left(-c\|h\|^{2\gamma}\right)$ e $C(u) = \left(a|u|^{2\alpha}+1\right)^{\delta}$ são as funções de covariância puramente espacial e puramente temporal, respectivamente.

3.2.1 Semivariograma Espaço–Temporal

O semivariograma espaço–temporal baseado no estimador de Matheron dado em (2.9) (ver p. 20) é obtido como (SCHABENBERGER; GOTWAY, 2005):

$$\hat{\gamma}(h,u) = \frac{1}{2|N(h,u)|} \sum_{N(h,u)} \{Z(s_i,t_i) - Z(s_j,t_j)\}^2 \tag{3.20}$$

onde $N(h,u)$ é o conjunto de pontos separados entre si por uma distância espacial igual a h e uma distância temporal igual a u e $|N(h,u)|$ é o número de pares distintos dentro deste conjunto.

O cálculo do semivariograma experimental em processos espaciais é de fundamental importância para a identificação do semivariograma teórico, porém na análise de processos espaço–temporais esta ferramenta nem sempre é utilizada. Na análise de processos dessa natureza não é possível fazer esta identificação direta de modelo, a menos que a função de covariância seja separável. Neste caso, poderíamos fazer o variograma (ou covariograma) para a parte espacial e outro para a parte temporal, identificar os modelos, transformá-los para as funções de covariância, e adicioná-los ou multiplicá-los.

3.2.2 Estimação dos Parâmetros do Modelo Espaço-Temporal

Sob a suposição de estacionariedade e gaussianidade os parâmetros do modelo podem ser estimados pelo método de máxima verossimilhança. Seja $Z(t) = (Z(s_1,t),...,Z(s_k,t))'$, $t = 1,...,n$, os dados observados em k localizações em n pontos no tempo. O logaritmo da função de verossimilhança de θ, onde θ são os parâmetros do modelo, baseado na matriz de dados $Z = \left(Z(1)',...,Z(n)'\right)$ pode ser escrito como (HUANG et al., 2007):

$$L(\theta, Z) = \log(2\pi)^{-kn/2} - \frac{1}{2}\left(\log|\Sigma| + Z'\Sigma^{-1}Z\right) \tag{3.21}$$

onde $\Sigma \equiv Var(Z)$ e " ′ " significa vetor transposto.

A seção seguinte apresenta o modelo geoestatístico proposto por Høst et al. (1995).

3.3 Modelo Geoestatístico proposto por Høst et al. (1995)

Considere um campo aleatório espaço–temporal $Z(s,t)$ como definido em (3.2) (ver p. 35), onde s é a localização espacial e t é o instante discreto do tempo no qual a observação é medida. O modelo espaço–temporal proposto por Høst et al. (1995) decompõe $Z(s,t)$ em três funções aleatórias espaço–temporais:

$$Z(s,t) = M(s,t) + S(s,t)R(s,t), \quad \forall (s,t) \in D \times T \tag{3.22}$$

sendo $M(s,t)$ a componente que representa a variação da média espacial, $S(s,t)$ a componente que modela o desvio-padrão da variável aleatória $Z(\cdot)$ e finalmente, $R(s,t)$ que é a componente espaço–temporal residual. A função aleatória $R(s,t)$ é temporalmente estacionária com média zero e variância igual a 1.

Os campos aleatórios que representam a média $(M(s,t))$ e o desvio-padrão $(S(s,t))$ ainda podem ser decompostos em fatores puramente espaciais e puramente temporais. A interação espaço–temporal é absorvida pela componente residual $R(s,t)$. A fatoração das componentes de média e de desvio padrão é dada como:

$$\begin{aligned} M(s,t) &= F(s) + \eta(t), \\ S(s,t) &= H(s)K(t) \end{aligned} \tag{3.23}$$

onde $F(s)$ e $\eta(t)$ modelam, respectivamente, o "efeito espacial" e o "efeito temporal" da função aleatória que representa a média, $H(s)$ modela o "efeito espacial" do desvio padrão e $K(t)$ é o fator de correção temporal de $S(s,t)$; $\eta(t)$ tem média igual à zero, e $K(t)$ tem média igual a 1.

A predição, que é frequentemente um dos objetivos finais da modelagem espaço–temporal, para um local não amostrado s_0 em um tempo específico $t_j, j = 1,2,...,n$ pode ser obtida como:

$$Z^*(s_0,t_j) = F^*(s_0) + \hat{\eta}(t_j) + H^*(s_0)\hat{K}(t_j)R^*(s_0,t_j), \quad \forall (s_0,t_j) \in D \times T \tag{3.24}$$

onde $F^*(s_0)$, $\hat{\eta}(t_j)$, $H^*(s_0)$, $\hat{K}(t_j)$ e $R^*(s_0,t_j)$ são as estimativas das componentes F, η, H, K e R para a localização não amostrada s_0.

A próxima seção apresenta os cálculos necessários para estimar as componentes do modelo dado em (3.24).

3.3.1 Estimação dos Parâmetros do Modelo de Høst et al. (1995)

Suponha que se tenha uma amostra de $Z(\cdot)$ constituída de k localizações s_i, $i = 1,2,...,k$, e n tempos t_j, $j = 1,2,...,n$. Os passos para estimar as componentes do modelo dado em (3.24) são apresentados a seguir (HØST et al., 1995):

Passo 1) Para cada localização fixa amostrada s_i, $i = 1,2,...,k$, calcular a média da série temporal respectiva, ou seja,

$$\hat{F}(s_i) = \frac{1}{n}\sum_{j=1}^{n} Z(s_i, t_j) \tag{3.25}$$

Assim, teremos o conjunto espacial com *k* informações de médias.

Passo 2) Calcular o variograma experimental da variável regionalizada[5] $\hat{F}(s)$. Identificar o modelo teórico e estimar os parâmetros. Fazer a krigagem ordinária no ponto s_0 e armazenar os pesos $\lambda^F(s_i)$, $i = 1,2,...,k$.

Passo 3) A primeira componente do modelo $F^*(\cdot)$ para a localização não amostrada s_0 é estimada como:

$$F^*(s_0) = \sum_{i=1}^{k} \lambda^F(s_i)\hat{F}(s_i) \tag{3.26}$$

onde $\lambda^F(s_i)$ são os pesos obtidos na krigagem ordinária no passo (2).

Passo 4) A componente $\hat{\eta}(t_j)$, para cada tempo $t_j, j = 1,2,...,n$, é calculada como:

[5] O termo *variável regionalizada* significa que o valor da característica medida está de alguma forma relacionada à sua disposição espacial (JOURNEL; HUIJBREGTS, 1978).

$$\hat{\eta}(t_j) = \sum_{i=1}^{k} \lambda^F(s_i)(Z(s_i, t_j) - \hat{F}(s_i)) \tag{3.27}$$

ou seja, $\hat{\eta}(t_j)$ é obtida pela diferença entre os valores reais observados de $Z(\cdot)$ e $\hat{F}(\cdot)$ para cada localização, ponderada pelos pesos obtidos na krigagem no passo (2).

Passo 5) A componente $\hat{H}^2(s_i)$ é dada por:

$$\hat{H}^2(s_i) = \frac{\sum_{j=1}^{n} \left[(Z(s_i, t_j) - \hat{F}(s_i) - \hat{\eta}(t_j) \right]^2}{n} \tag{3.28}$$

Passo 6) Calcular o variograma experimental da variável regionalizada $\hat{H}(s)$. Identificar o modelo teórico e estimar os parâmetros. Fazer a krigagem ordinária no ponto s_0 e armazenar os pesos $\lambda^H(s_i)$, $i = 1, 2, ..., k$.

Passo 7) A componente $H^*(\cdot)$ para a localização s_0 é estimada por:

$$H^*(s_0) = \sum_{i=1}^{k} \lambda^H(s_i)\hat{H}(s_i) \tag{3.29}$$

onde $\lambda^H(s_i)$ são os pesos obtidos na krigagem ordinária no passo (6) para cada localização s_i, $i = 1, 2, ..., k$.

Passo 8) Cálculo da constante $\hat{v}^2$, onde $v^2 = \{E[H(s)]\}^2$:

$$\hat{v}^2 = \sum_{i=1}^{k} \lambda^F(s_i)\hat{H}^2(s_i) \tag{3.30}$$

onde $\lambda^F(s_i)$ são os pesos originados da krigagem ordinária calculados no passo (2).

Passo 9) A componente $\hat{K}^2(t_j)$ é estimada como:

$$\hat{K}^2(t_j) = \frac{\sum_{i=1}^{k} \lambda^F(s_i)\left[Z(s_i,t_j) - \hat{F}(s_i) - \hat{\eta}(t_j)\right]^2}{\hat{v}^2} \tag{3.31}$$

onde $\lambda^F(s_i)$ também são os pesos obtidos na krigagem ordinária da variável regionalizada $\hat{F}(s_i)$, $i = 1,2,...,k$.

Passo 10) A componente espaço-temporal residual é obtida como:

$$\hat{R}(s_i,t_j) = \frac{Z(s_i,t_j) - \hat{F}(s_i) - \hat{\eta}(t_j)}{\hat{H}(s_i)\hat{K}(t_j)} \tag{3.32}$$

Passo 11) Calcular o variograma experimental da variável regionalizada $\hat{R}(s_i,t_j)$, identificar o modelo teórico e estimar os parâmetros. Fazer a krigagem simples no ponto s_0 e armazenar os pesos $\lambda^R(s,t)$.

Passo 12) Calcular a componente residual estimada em s_0:

$$R^*(s_0,t_j) = \sum_{i=1}^{k} \lambda^R(s_i,t_j)\hat{R}(s_i,t_j) \tag{3.33}$$

Passo 13) Somando as componentes obtidas nos passos 3, 4, 7, 9 e 12 obtemos a previsão da característica de interesse no local s_0 para o tempo $t_j, j = 1,2,...,n$ que é dada por:

$$Z^*(s_0,t_j) = F^*(s_0) + \hat{\eta}(t_j) + H^*(s_0)\hat{K}(t_j)R^*(s_0,t_j) \tag{3.34}$$

3.3.2 Proposta de Kyriakidis e Journel (1999)

A estimação dos parâmetros do modelo (3.24) sugerida por Kyriakidis e Journel (1999) é mais simples se comparada com o procedimento anterior, pois não precisamos calcular os vetores de pesos das componentes $\hat{H}(s_i)$ e $\hat{R}(s_i,t_j)$. Os autores utilizam o vetor de pesos originado da krigagem ordinária da componente $\hat{F}(s_i)$, calculada no passo (2) do ajuste

anterior, para ponderar as estimativas nas equações (3.29) e (3.33). Logo, não precisamos realizar os passos (6) e (11), e as etapas para a estimação das componentes são as seguintes:

1) Devemos estimar as componentes $\hat{F}(s_i), \hat{\eta}(t_j), \hat{H}(s_i), \hat{v}^2, \hat{K}(t_j)$ e $\hat{R}(s_i, t_j)$ como foi mostrado nos passos 1, 4, 5, 8, 9 e 10 respectivamente do algoritmo anterior.

2) As componentes estimadas do modelo (3.24) são dadas por:

$$\begin{cases} F^*(s_0) = \sum_{i=1}^{k} \lambda^F(s_i)\hat{F}(s_i) \\ H^*(s_0) = \sum_{i=1}^{k} \lambda^F(s_i)\hat{H}(s_i) \\ R^*(s_0, t_j) = \sum_{i=1}^{k} \lambda^F(s_i)\hat{R}(s_i, t_j) \end{cases} \tag{3.35}$$

onde $\lambda^F(s_i)$, $i = 1, 2, ..., k$, são os pesos obtidos com a krigagem da componente *F* calculada no passo (1).

Observamos então que os pesos das componentes $H^*(s_0)$ e $R^*(s_0, t_j)$ é o que diferencia este processo daquele método de estimação proposto por Høst et al. (1995). Na proposta de Kyriakidis e Journel (1999) todos os pesos são provenientes da krigagem ordinária da variável regionalizada $\hat{F}(s_i)$.

Essa metodologia, ao contrário da família de funções de covariância proposta por Gneiting (2002) e apresentada na seção 3.2, somente permite previsão de Z para uma localidade s_0 nos tempos que foram avaliados na amostra. Logo, se pretendemos estimar Z em uma localidade qualquer em um determinado tempo t, é necessário termos informações da vizinhança deste ponto a ser predito no mesmo instante de tempo t, pois caso contrário, esta previsão não pode ser obtida (ou calculada).

A seção subseqüente mostra o modelo de séries temporais proposto por Niu et al. (2003).

3.4 Modelo de Séries Temporais proposto por Niu et al. (2003)

Esta seção mostra uma classe de modelos de série temporal sazonal espacial proposta por Niu et al. (2003) para sistemas do tipo lattice[6]. Os modelos apresentados são modelos desenvolvidos inicialmente para séries temporais e que incluem a componente espacial, ou seja, leva-se em consideração na modelagem a distribuição espacial dos pontos amostrados. A seguir fazemos uma breve descrição desta classe de modelos e com o intuito de facilitar o entendimento mostramos o exemplo de aplicação analisado no artigo de Niu et al. (2003).

Seja $\{Y_{ij}(t),\ i=1,2,...,m;\ j=1,2,...,n;\ t=1,2,...,T\}$ o processo espaço–temporal onde (i,j) determina o local de coleta dos dados. Temos que i é a latitude e j é a longitude. O sistema *lattice* pode ser esquematizado como na Figura 3.1:

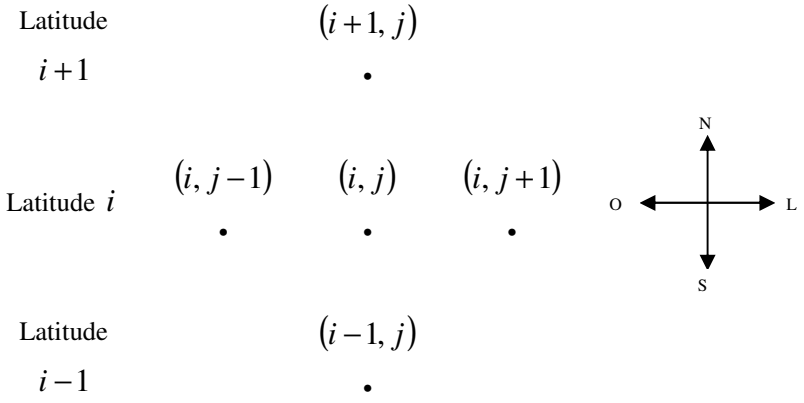


Figura 3.1. Representação esquemática do sistema *lattice*.

A vizinhança espacial de ordem *r* é definida como:

$$N_{ij}(r)=\left\{(a,b):(a,b)\in L_{m\times n};0\le\sqrt{(a-i)^2+(b-j)^2}\le r\right\} \tag{3.36}$$

onde $L_{m\times n}$ é o domínio espacial. A vizinhança espacial de ordem $r=1$ diz que o valor de um campo aleatório em um dado ponto (i,j) é influenciado, diretamente, somente pelos vizinhos mais próximos e a sua localização não é relevante.

O modelo espaço-temporal $\{Y_{ij}(t)\}$ proposto por Niu et al. (2003) é dado por:

[6] *Lattice* é o sistema o qual as estações dispostas em um grid fixo são identificadas, ou localizadas, a partir das coordenadas de latitude e de longitude.

$$\left(1-B^{s}\right)^{D}(1-B)^{d}\,Y_{ij}(t)=\xi_{ij}(t) \tag{3.37}$$

onde B é o operador de retardo temporal que é obtido como: $BY_{ij}(t)=Y_{ij}(t-1)$; S é o período sazonal, e D e d são, respectivamente, os graus de diferenciação sazonal e de tendência.

O modelo em (3.38) fornece a estrutura do ruído do processo $\{\xi_{ij}(t)\}$ mostrado em (3.37), após a diferenciação:

$$\xi_{ij}(t)=\sum_{k=0}^{p}\sum_{(a,b)\in N_{ij(r)}}\beta_{a\,b\,k}\;\xi_{i-a,\,j-b}(t-k)+\sum_{k=1}^{p}\phi_{ijk}\,\xi_{ij}(t-k)+\varepsilon_{ij}(t)-\sum_{l=1}^{q}\theta_{ijl}\,\varepsilon_{ij}(t-l) \tag{3.38}$$

Na direção temporal este modelo se reduz a um ARMA de ordem p e ϕ_{ijk} são os parâmetros da parte autoregressiva (AR), e de ordem q e θ_{ijl} são os parâmetros da parte média móvel (MA). A variabilidade espacial no modelo é incorporada pela dependência das coordenadas de latitude e de longitude. Os parâmetros β_{abk} estão relacionados com a vizinhança espacial. Temos que os $\varepsilon_{ij}(t)$`s são independentes e identicamente distribuídos por uma normal com média igual a zero e variância igual a σ^2. O modelo (3.38) é denotado por modelo espaço–temporal autoregressivo e de média móvel (NIU et al., 2003). Para outras informações acerca deste modelo, tais como, funções de covariância, estimação de parâmetros e predição deve-se consultar o trabalho de Niu et al. (2003).

Outros modelos podem ser obtidos a partir da simplificação do modelo em (3.38). Se a vizinhança espacial é de ordem igual a 1, i.e. $r=1$, o modelo em (3.38) pode ser reescrito como em (3.39):

$$\begin{aligned}\xi_{ij}(t) \;&=\sum_{k=0}^{p}\left[\beta_{k1}\,\xi_{i-1,\,j}(t-k)+\beta_{k2}\,\xi_{i+1,\,j}(t-k)+\alpha_{k1}\,\xi_{i,\,j-1}(t-k)+\alpha_{k2}\,\xi_{i,\,j+1}(t-k)\right]+\\ &+\sum_{k=1}^{p}\phi_{ijk}\xi_{ij}(t-k)+\varepsilon_{ij}(t)-\sum_{l=1}^{q}\theta_{ijl}\,\varepsilon_{ij}(t-l)\end{aligned} \tag{3.39}$$

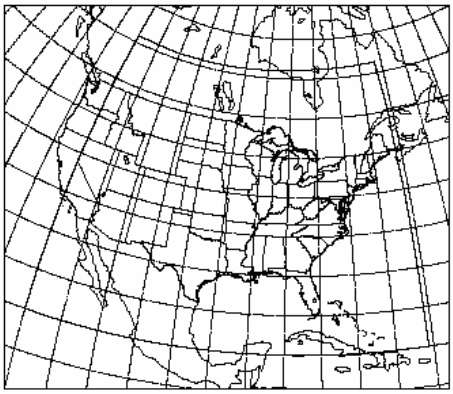
onde os parâmetros β e α estão relacionados respectivamente com a vizinhança espacial na direção norte-sul (latitude) e na direção leste-oeste (longitude).

Os modelos espaço-temporais autoregressivos e de médias móveis propostos por Niu et al. (2003) somente permitem a previsão temporal de Z nas localizações observadas na amostra. Para predizer uma observação em um tempo $t, t>T$ em uma localização qualquer s, precisamos das informações da característica de interesse na própria localização s e nos

vizinhos próximos em tempos anteriores. A seção seguinte mostra o exemplo de aplicação apresentado no artigo de Niu et al. (2003).

3.4.1 Exemplo de Aplicação

Como ilustração, mostramos o modelo de série temporal sazonal espacial aplicado a um conjunto de dados referente a *alturas geopotenciais* apresentado em Niu et al. (2003). De acordo com Byers[7] (1974) apud Niu et al. (2003) a *altura geopotencial* é uma medida da densidade do ar. Esta medida é utilizada para predizer o clima a médio (6 dias a 2 semanas) e a longo prazo (comportamentos mensais ou sazonais). Estamos analisando uma série de 516 valores mensais amostrada no período entre janeiro de 1946 à dezembro de 1988. As localizações estão dispostas em um lattice de tamanho 10×10, ou seja, temos um vetor de observações em cada uma das 100 estações de monitoramento. O lattice é delimitado pela latitude e pela longitude que variam, respectivamente, de 20ºN a 56ºN e de 66ºO a 120ºO. A região de estudo é mostrada na Figura 3.2.



Fonte: Niu, Mckeague e Elsner (2003).

Figura 3.2. Região de estudo.

Neste estudo diversos modelos foram construídos pelos autores para tentar explicar a variabilidade espaço-temporal dos dados. A identificação do melhor modelo foi feita avaliando a habilidade preditiva do mesmo, e nesta análise, foi suposto que as doze últimas observações da série eram desconhecidas, sendo estas estimadas (ou preditas) pelo modelo ajustado.

[7] BYERS, H. R. **General Meteorology**. New York: McGraw-Hill, 1974.

Niu et al. (2003) ajustam quatro modelos de série temporal sazonal (com sazonalidade igual a 12) espacial aos dados. Estes modelos são especificados a seguir:

i. Modelo 1. Modelo espaço-temporal simétrico de ordem 1.

$$\begin{aligned}\xi_{ij}(t) &= \beta\left[\xi_{i-1,j}(t)+\xi_{i+1,j}(t)\right]+\alpha\left[\xi_{i,j-1}(t)+\xi_{i,j+1}(t)\right]+\\ &\quad \phi_i\xi_{ij}(t-1)+\varepsilon_{ij}(t)-\theta_i\varepsilon_{ij}(t-12)\end{aligned} \tag{3.40}$$

ii. Modelo 2. Modelo espaço-temporal assimétrico de ordem 1.

$$\begin{aligned}\xi_{ij}(t) &= \beta_{01}\xi_{i-1,j}(t)+\beta_{02}\xi_{i+1,j}(t)+\alpha_{01}\xi_{i,j-1}(t)+\alpha_{02}\xi_{i,j+1}(t)+\\ &\quad \phi_i\xi_{ij}(t-1)+\varepsilon_{ij}(t)-\theta_i\varepsilon_{ij}(t-12)\end{aligned} \tag{3.41}$$

iii. Modelo 3. Modelo espacial de ordem 1 e temporal de lag 1.

$$\begin{aligned}\xi_{ij}(t) &= \beta_{11}\xi_{i-1,j}(t-1)+\beta_{12}\xi_{i+1,j}(t-1)+\alpha_{11}\xi_{i,j-1}(t-1)+\alpha_{12}\xi_{i,j+1}(t-1)+\\ &\quad \phi_i\xi_{ij}(t-1)+\varepsilon_{ij}(t)-\theta_i\varepsilon_{ij}(t-12)\end{aligned} \tag{3.42}$$

iv. Modelo 4. Modelo espacial de ordem 2 e temporal de lag 1.

$$\begin{aligned}\xi_{ij}(t) &= \beta_{11}\xi_{i-1,j}(t-1)+\beta_{12}\xi_{i+1,j}(t-1)+\alpha_{11}\xi_{i,j-1}(t-1)+\alpha_{12}\xi_{i,j+1}(t-1)+\\ &\quad \beta_{13}\xi_{i-1,j-1}(t-1)+\beta_{14}\xi_{i+1,j-1}(t-1)+\alpha_{13}\xi_{i-1,j+1}(t-1)+\alpha_{14}\xi_{i+1,j+1}(t-1)+\\ &\quad \phi_i\xi_{ij}(t-1)+\varepsilon_{ij}(t)-\theta_i\varepsilon_{ij}(t-12)\end{aligned} \tag{3.43}$$

Segundo Niu et al. (2003) o modelo apresentado em (3.41) fornece o melhor ajuste aos dados em termos da maximização da função de verossimilhança condicional se comparado com os modelos (3.40), (3.42) e (3.43). A soma de quadrados das diferenças entre o valor real e o valor predito, ou seja, a soma de quadrados dos erros foi calculada usando os modelos (3.42) e (3.43). Este cálculo utiliza as doze observações que foram desconsideradas na amostra. O modelo (3.43) foi o que apresentou o menor valor. Segundo os autores, devido a problemas de inversão de matrizes, os modelos (3.40) e (3.41) foram os que apresentaram os piores valores. Os autores concluem dizendo que o modelo (3.41) é preferível em relação aos outros para explicar a estrutura dos dados, porém a habilidade preditiva deste modelo não é boa.

Nas seções seguintes sugerimos algumas modificações nos modelos de Høst et al. (1995) e de Niu et al. (2003) para a análise de dados espaço-temporais, e além disso propomos uma estratégia de análise que combina esses dois modelos de geoestatística e de séries temporais em duas etapas. Estes modelos serão ajustados aos dados apresentados nos Capítulos 4, 5 e 6.

3.5 Modificações nos Modelos de Høst et al. (1995) e de Niu et al. (2003)

Os modelos espaço-temporais buscam descrever probabilisticamente a variabilidade dos dados no espaço e no tempo e o interesse primordial da análise espaço-temporal consiste na estimação dos parâmetros do modelo e/ou na predição da característica de interesse. Podem-se observar três situações distintas de predição: a primeira considera a predição espacial (ou interpolação espacial) em tempos observados na amostra, a segunda consiste na predição temporal em localizações amostradas e a terceira, faz-se a predição em locais e tempos que não foram amostrados. O tipo de predição está associado com as informações (ou dados) disponíveis e com o objetivo do estudo.

Para as duas primeiras situações de predição sugerimos algumas modificações nos modelos propostos por Høst et al. (1995) e por Niu et al. (2003) que são descritas a seguir. No terceiro caso de predição propomos uma estratégia de modelagem que combina esses modelos modificados em duas etapas distintas e esta proposta é discutida na seção 3.6.

3.5.1 Modificações no Modelo de Høst et al. (1995)

Quando temos interesse em predizer a característica de interesse em novas localidades considerando os mesmos tempos que foram amostrados, sugerimos utilizar o modelo de Høst et al. (1995) com as componentes estimadas da forma mostrada em Kyriakidis e Journel (1999) e propomos uma modificação na quantidade de vizinhos utilizada na predição das observações.

Sabe-se pela dependência espacial que os vizinhos mais próximos do ponto a ser predito são mais importantes e têm maior influência que aqueles pontos mais afastados, ou seja, supõe-se que as observações mais próximas geograficamente tenham um comportamento mais semelhante entre si do que aquelas separadas por distâncias maiores. Desta forma, ao invés de utilizar todos os vizinhos (ou todos os valores amostrais) no respectivo tempo t da amostra para estimar a nova observação como proposto por Kyriakidis e Journel (1999), sugerimos variar esta quantidade de vizinhos. Espera-se que somente a informação dos pontos mais próximos seja relevante resultando em um menor número de vizinhos na análise.

Nesta dissertação a quantidade de vizinhos utilizada na análise é calculada de duas formas distintas. A primeira é baseada no menor valor do erro quadrático médio (EQM), i.e., na soma de quadrados da diferença entre os valores observados e os valores preditos em todas as localizações e tempos em que se deseja fazer a predição. A segunda forma considera a

maior distância, denotada por *di,* dentre a qual existe dependência espacial entre as observações. O método pelo EQM não é realista, pois é necessário conhecer os valores reais das observações nos locais e tempos de predição para seu cálculo e nesta dissertação este método será utilizado apenas para efeito de comparação de modelos e para a avaliação do método da distância *di*. No ajuste dos modelos aos dados reais mostrados nos capítulos subseqüentes fornecemos mais detalhes sobre a escolha do número de vizinhos para a predição.

A variação do número de vizinhos da proposta de Kyriakidis e Journel (1999) ocasiona algumas modificações nos cálculos de predição. Se a quantidade de vizinhos for igual a 1, então a componente $\hat{H}_i^2$, $i = 1,2,...,k$, calculada no passo (5) (ver seção 3.3.1, p. 46) é igual à zero, e conseqüentemente as próximas componentes (v, K, R) não são estimadas. Portanto, a predição de uma determinada localidade quando utilizamos apenas o vizinho mais próximo é igual à soma das componentes F e η (ver seção 3.3.1, p. 45). Outra alteração é no cálculo da componente $\hat{K}_j^2$, $j = 1,2,...,n$, mostrada no passo (9) (ver seção 3.3.1, p. 46), no qual decidimos tomar o módulo do valor obtido, visto que esta componente não pode ter valores negativos[8].

3.5.2 Modificações no Modelo de Niu et al. (2003)

Quando temos o interesse em predizer observações para tempos futuros nas localizações que compõem a amostra, construímos um modelo baseado nas idéias apresentadas em Niu et al. (2003). Suponha que se tenha uma amostra de $Z(\cdot)$ constituída de k localizações s_i, $i = 1,2,...,k$, e n tempos t_j, $j = 1,2,...,n$. O modelo trabalhado nesta dissertação pode ser escrito como em (3.44):

$$\begin{aligned} Z(s,t) = \beta_0 &+ \beta_1 Z(s,t-1) + \beta_2 Z(v_1,t-1) + \beta_3 Z(v_2,t-1) + ... \\ &+ \beta_k Z(v_{k-1},t-1) + \varepsilon(s,t) \end{aligned} \tag{3.44}$$

onde v_i, $i = 1,2,...,k-1$, é o i-ésimo vizinho mais próximo e ε é o erro associado ao modelo. Os vizinhos mais próximos são aqueles que apresentam a menor distância Euclidiana do ponto a ser predito. Uma das diferenças do modelo dado em (3.44) e o proposto por Niu et al.

[8] Para o cálculo da componente residual espaço-temporal, utilizamos a raiz quadrada desta componente.

(2003) é que neste caso não supomos que as observações estão dispostas em um *lattice*, i.e., em um *grid* fixo ou uma malha regular (ver Figura 3.2, p. 49).

O modelo (3.44) diz que a característica de interesse medida em uma determinada localidade s no instante de tempo t depende do valor desta característica na própria localização e nos vizinhos no tempo imediatamente anterior $(t-1)$, adicionada de uma constante β_0. A parte temporal do modelo é um autoregressivo de ordem 1 (AR(1)), dessa forma o coeficiente β_1 esta sujeito à restrição: $-1 < \beta_1 < 1$.

Os coeficientes do modelo, $\beta_0, \beta_1, ..., \beta_k$, são estimados pelo método de mínimos quadrados ordinários, ou seja, estes valores são obtidos pela minimização do quadrado da soma de resíduos que na forma matricial é escrita como:

$$\text{SQE} = \left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right) \tag{3.45}$$

onde,

$$Y = \begin{bmatrix} Z(s_1,t) \\ Z(s_1,t-1) \\ \vdots \\ Z(s_1,2) \\ \vdots \\ Z(s_k,t) \\ Z(s_k,t-1) \\ \vdots \\ Z(s_k,2) \end{bmatrix}, X = \begin{bmatrix} 1 & Z(s_1,t-1) & Z(v_1,t-1) & Z(v_2,t-1) & \cdots & Z(v_{k-1},t-1) \\ 1 & Z(s_1,t-2) & Z(v_1,t-2) & Z(v_2,t-2) & \cdots & Z(v_{k-1},t-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z(s_1,1) & Z(v_1,1) & Z(v_2,1) & \cdots & Z(v_{k-1},1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z(s_k,t-1) & Z(v_1,t-1) & Z(v_2,t-1) & \cdots & Z(v_{k-1},t-1) \\ 1 & Z(s_k,t-2) & Z(v_1,t-2) & Z(v_2,t-2) & \cdots & Z(v_{k-1},t-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z(s_k,1) & Z(v_1,1) & Z(v_2,1) & \cdots & Z(v_{k-1},1) \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}.$$

O vetor Y tem tamanho igual a $(k \times (t-1))$, a matriz X tem dimensão igual a $(k \times (t-1)) \times (k+1)$ e β é um vetor de tamanho igual a $k+1$. Niu et al. (2003) estimam as componentes do modelo usando a função de máxima verossimilhança condicional, pois segundo os autores (NIU et al., 2003, p. 120) "o método de máxima verossimilhança envolve cálculos computacionais pesados".

A média da soma de quadrados do erro (MSQE) é dada por:

$$\text{MSQE} = \frac{\text{SQE}}{k \times n} \tag{3.46}$$

onde *k* é o número de localizações e *n* é a quantidade de tempos utilizada na estimação dos parâmetros do modelo.

3.6 Estratégia de Análise de Dados: Combinação de Modelos de Geoestatística e de Séries Temporais

Nesta seção propomos uma estratégia de modelagem de dados espaço-temporais que combina os modelos de Høst et al (1995) e de Niu et al. (2003) com as modificações descritas previamente em duas etapas de análise. Esta proposta tem por objetivo permitir (ou habilitar) a predição da característica de interesse em localizações e instantes de tempo não observados na amostra, sendo essa estratégia uma alternativa prática às funções de covariância de Gneiting (2002).

A primeira etapa deste procedimento (da estratégia de análise de dados) consiste na predição das observações nas localizações não amostradas, porém em tempos observados na amostra. Esta predição é realizada ajustando-se o modelo de Høst et al. (1995) aos dados com as modificações sugeridas previamente em 3.5.1. Na segunda etapa atualizamos o banco de dados com essas predições e aplicamos o modelo apresentado na equação (3.44) baseado nas idéias de Niu et al. (2003), para estimar as observações nestas mesmas localidades para tempos futuros.

Uma das motivações em se construir esta modelagem foi a possibilidade de implementação computacional, visto que há uma carência em programas para a análise de dados indexados no espaço e no tempo. As metodologias de Høst et al. (1995) e de Kyriakidis e Journel (1999), embora muito mencionadas na literatura, não aparecem implementadas em trabalhos práticos publicados. O pacote *RandomFields* implementado no *software* R permite a análise e a simulação de processos espaço – temporais, porém segundo Silva (2006, p. 30) "a modelagem espaço-temporal no pacote "*RandomFields*" está em desenvolvimento e por conta disso seus procedimentos ainda não são exaustivos e claramente documentados e existem poucos trabalhos que se utilizam deste pacote".

Nos capítulos subseqüentes avaliamos estas propostas (modelos modificados de Høst et al. (1995) e de Niu et al. (2003), e a combinação desses modelos) aplicando estas modelagens a dados reais distintos. O primeiro banco é referente a taxa de criminalidade no estado de Minas Gerais, o segundo é relativo a armazenagem de água em um solo cultivado com citros e o terceiro é concernente a incidência de Aids nas microrregiões de Minas Gerais.

Os dados também são ajustados por funções de covariância separável e não-separável da família de Gneiting (2002) e os resultados das predições são comparados. A viabilidade do ajuste dos modelos da classe de Gneiting (2002) aos dados desta dissertação é devido ao estudo de simulação de campos aleatórios espaço-temporais feito por Silva (2006) cujo

objetivo era aperfeiçoar a compreensão das funcionalidades do pacote *RandomFields* (2006), além de identificar os modelos implementados no programa. Silva (2006) ainda apresenta dois exemplos de aplicação que também auxiliaram no ajuste destas funções de covariância aos dados tratados nesta dissertação.

Capítulo 4

Estudo de Caso: Taxa de Criminalidade no Estado de Minas Gerais

Neste capítulo ajustamos os modelos apresentados no capítulo 3 aos dados da taxa de criminalidade no estado de Minas Gerais. Estes modelos se referem à proposta de Høst et al. (1995) com as modificações apresentadas previamente (ver seção 3.5.1, p. 53), o modelo baseado nas idéias de Niu et al. (2003) (ver seção 3.5.2, p. 54), a combinação destes modelos (ver seção 3.6, p. 56) e o ajuste dos dados por funções de covariância espaço-temporal separável e não-separável da família de Gneiting (2002) (ver seção 3.2, p. 40).

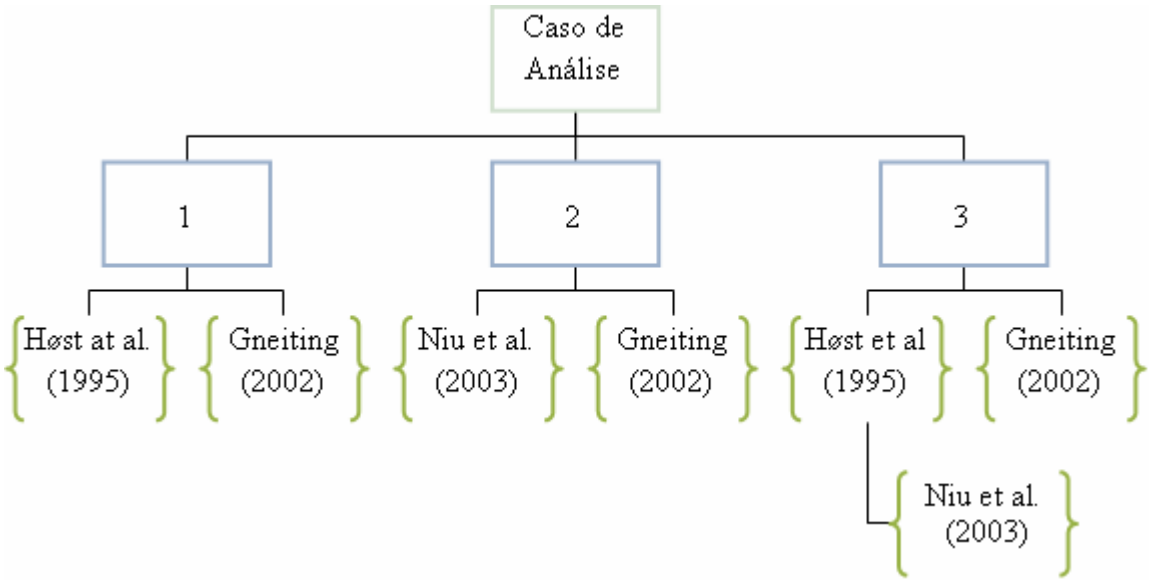
4.1 Introdução

Os dados apresentados neste capítulo e nos capítulos subseqüentes (Capítulos 5 e 6) foram analisados nesta dissertação de três formas distintas. As formas de análise estão relacionadas com o objetivo do estudo e com as informações (ou dados) disponíveis. Por simplificação utilizaremos a seguinte notação ao longo da dissertação: Caso 1 se refere aos modelos cujo objetivo é a predição espacial em tempos observados na amostra; Caso 2 é concernente a modelos que fazem à predição temporal em localizações amostradas; e finalmente no Caso 3 têm-se modelos cuja predição é calculada para locais e tempos não observados na amostra.

No Caso 1 aplicou-se o modelo proposto por Høst et al. (1995) com as componentes estimadas pelo método apresentado em Kyriakidis e Journel (1999) e com as modificações descritas na seção 3.5.1 (ver p. 53). O modelo baseado na proposta de Niu et al. (2003) (ver seção 3.5.2, p. 54) foi utilizado no Caso 2 de análise. E finalmente, no Caso 3 os dois modelos anteriores foram combinados em 2 etapas, de acordo com o que foi descrito na seção 3.6 (ver p. 56).

As funções de covariância espaço-temporal separáveis e não-separáveis da família de Gneiting (2002) também foram ajustadas aos dados nas três situações distintas de predição: Casos 1, 2 e 3. Os resultados foram comparados com aqueles obtidos pelo ajuste dos modelos de Høst et al. (1995) e de Niu et al. (2003), e pela combinação de ambos.

O fluxograma a seguir apresenta os modelos ajustados em cada um dos casos de análise.



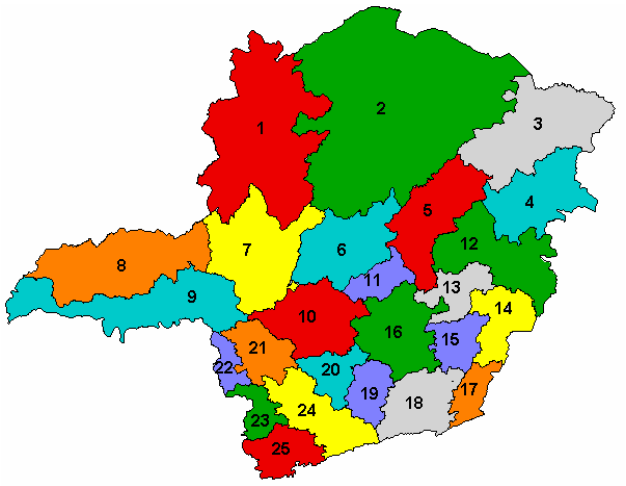
Fluxograma 4.1. Modelos ajustados em cada caso de análise.

A adequação do ajuste dos modelos aos dados foi feita pela predição espacial e/ou temporal da característica de interesse em locais e/ou tempos que foram considerados desconhecidos nas análises de dados, ou seja, algumas observações escolhidas aleatoriamente foram separadas da base de dados original e supusemos não conhecer o seu valor com o objetivo de validar o modelo ajustado.

A seção seguinte apresenta uma descrição dos dados analisados neste capítulo.

4.2 Descrição dos Dados

Os dados analisados neste capítulo se referem à taxa bruta de criminalidade violenta por 100.000 habitantes nas 25 regiões administrativas do estado de Minas Gerais entre os anos de 1986 e 1997. A localização de cada região é feita pelas coordenadas de latitude e longitude do município sede da região administrativa. Os dados foram obtidos no site: http://www.nadd.prp.usp.br/cis/index.aspx. As regiões administrativas são apresentadas na Figura 4.1.



1. Noroeste
2. Norte de Minas
3. Vale do Jequitinhonha
4. Vale do Mucuri
5. Alto do Jequitinhonha
6. Médio São Francisco
7. Alto Paranaíba
8. Vale do Paranaíba
9. Vale do Rio Grande
10. Alto São Francisco
11. Alto Rio das Velhas
12. Vale do Rio Doce
13. Vale do Aço
14. Vertente do Caparaó
15. Vale do Rio Piranga
16. Central
17. Vale do Rio Pomba
18. Mata
19. Campo das Vertentes
20. Alto Rio Grande
21. Médio Rio Grande
22. Sudoeste
23. Alto Rio Pardo
24. Baixo Sapucaí
25. Vale do Sapucaí

Figura 4.1. Regiões administrativas do estado de Minas Gerais.

Segundo Beato F. et al. (1998) a criminalidade violenta engloba os crimes classificados pela Polícia Militar de Minas Gerais como: homicídio, homicídio tentado, estupro, roubo, roubo a mão armada, roubo de veículos, roubo de veículos a mão armada e seqüestro.

A Tabela 4.1 mostra a média, o desvio-padrão e a soma das taxas de criminalidade nas 25 regiões administrativas no período de 1986 a 1997. Esta tabela ainda inclui o município sede de cada região e a população correspondente no ano de 1996.

As regiões administrativas Central, do Vale do Rio Doce e Vale do Paranaíba são as que apresentam os valores mais altos da taxa de criminalidade no período de 1986 a 1997 e a variabilidade da taxa nestas regiões é alta. Estas regiões correspondem a algum dos municípios sede mais populosos do estado de Minas Gerais que são: Belo Horizonte, Governador Valadares e Uberlândia.

As regiões com menor índice de criminalidade são: Sudoeste, Alto Rio Grande e Baixo Sapucaí. Os municípios sede menos populosos que correspondem às regiões administrativas do Vale do Jequitinhonha, Alto do Jequitinhonha e Vale do Rio Piranga são respectivamente: Araçuaí, Diamantina e Ponte Nova.

Tabela 4.1 – Análise Descritiva da Taxa de Criminalidade por Região Administrativa.

Região Administrativa	Município Sede	Média	Desvio Padrão	Soma	População
Alto do Jequitinhonha	Diamantina	62,29	10,46	747,46	44.223
Alto Paranaíba	Patos de Minas	69,99	11,28	839,92	139.357
Alto Rio das Velhas	Sete Lagoas	113,79	23,53	1365,47	215.068
Alto Rio Grande	Lavras	39,06	5,92	468,76	88.290
Alto Rio Pardo	Poços de Caldas	52,52	8,38	630,18	154.474
Alto São Francisco	Divinópolis	51,33	4,88	615,96	207.981
Baixo Sapucaí	Varginha	41,58	9,28	499,01	124.501
Campo das Vertentes	São João del Rei	50,91	13,43	610,97	82.952
Central	Belo Horizonte	231,86	66,09	2782,31	2.399.920
Mata	Juiz de Fora	116,68	15,02	1400,12	509.126
Médio Rio Grande	Passos	50,47	6	605,65	106.516
Médio São Francisco	Curvelo	89,31	9	1071,67	73.791
Noroeste	Paracatu	117,92	21,83	1415,05	84.411
Norte de Minas	Montes Claros	79,17	11,05	950,02	348.990
Sudoeste	São Sebastião do Paraíso	37,25	9,85	446,95	65.197
Vale do Aço	Coronel Fabriciano	115,14	10,67	1381,71	104.851
Vale do Jequitinhonha	Araçuaí	61,2	4,19	734,43	37.109
Vale do Mucuri	Teófilo Otoni	127,86	21,97	1534,27	127.530
Vale do Paranaíba	Uberlândia	150,42	68,57	1805,08	600.367
Vale do Rio Doce	Governador Valadares	158,48	43,4	1901,81	259.407
Vale do Rio Grande	Uberaba	142,85	32,41	1714,21	285.093
Vale do Rio Piranga	Ponte Nova	59,57	8,49	714,82	57.344
Vale do Rio Pomba	Muriaé	64,78	14,32	777,31	100.068
Vale do Sapucaí	Pouso Alegre	52,27	12,63	627,18	125.206
Vertente do Caparaó	Caratinga	75,97	7,66	911,66	82.633

Para a validação dos modelos espaço-temporais ajustados, algumas observações foram separadas do banco de dados original e consideradas desconhecidas para efeito de análise. A qualidade do ajuste (ou adequação do modelo) foi avaliada pela predição espacial e/ou temporal destas observações. Estas observações escolhidas aleatoriamente na base de dados correspondem a cinco regiões administrativas (ou 20% dos dados): Alto do Jequitinhonha, Central, Noroeste, Vale do Rio Grande e Vale do Sapucaí. A localização destas regiões é mostrada na Figura 4.2.

Figura 4.2. Regiões consideradas desconhecidas na análise.

A Figura 4.3 mostra o comportamento da taxa de criminalidade nestas cinco regiões de validação nos anos de 1986 a 1997 conjuntamente com os quatro vizinhos mais próximos.

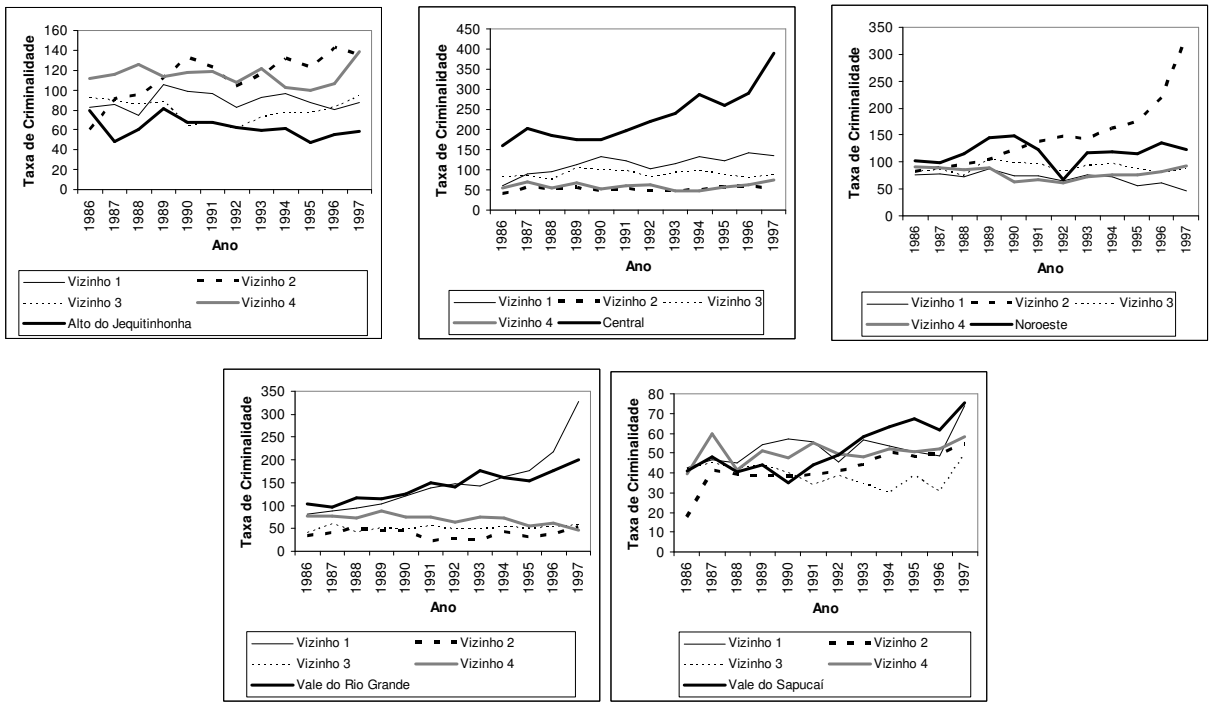


Figura 4.3. Evolução da taxa de criminalidade nas cinco regiões de validação e nos vizinhos mais próximos.

Observamos pela Figura 4.3 que a taxa de criminalidade nas regiões administrativas do Vale do Rio Grande, Vale do Sapucaí e Noroeste são semelhantes com os valores da taxa nos vizinhos mais próximos. A região Central apresenta valores superiores da taxa de criminalidade se comparada com os seus vizinhos.

No Caso 1 de análise o objetivo é predizer a taxa de criminalidade nestas 5 regiões consideradas desconhecidas em todos os tempos observados na amostra. Sendo assim trabalhamos com um banco de dados composto por 20 regiões administrativas e para cada região tem-se a informação da taxa de criminalidade nos anos de 1986 a 1997.

O propósito do Caso 2 de análise é predizer a taxa nestas mesmas regiões de validação nos três últimos anos, i.e., de 1995 a 1997. A escolha de 3 tempos para a predição da taxa se deve ao pequeno número de observações da série temporal (12 instantes de tempo). Neste caso o banco de dados utilizado é formado pelas 25 regiões administrativas, porém a característica de interesse é observada somente no período de 1986 a 1994, i.e., supomos que não conhecemos a taxa de criminalidade nas 25 regiões nos anos de 1995, 1996 e 1997.

No Caso 3 temos informações de 20 regiões em 9 instantes de tempo (de 1986 a 1994) e a qualidade de ajuste do modelo é avaliada pela predição da taxa de criminalidade nas 5

regiões nos anos de 1994 a 1997. A diferença do Caso 3 de análise para o Caso 2 é a quantidade de regiões que compõem a base de dados. Consideramos que a taxa de criminalidade nos anos de 1995, 1996 e 1997 não é conhecida para nenhuma das 25 regiões administrativas do estado de Minas Gerais. Além disso, supomos também não conhecer a taxa nos anos de 1986 a 1994 nas cinco regiões utilizadas para a avaliação do ajuste do modelo.

A Tabela 4.2 mostra os modelos ajustados a cada um dos casos e as informações (quantidade de localizações e tempos) disponíveis na base de dados para a predição. Os resultados dos modelos ajustados aos Casos 1, 2 e 3 de análise são apresentados nas próximas seções: 4.3 a 4.5 respectivamente.

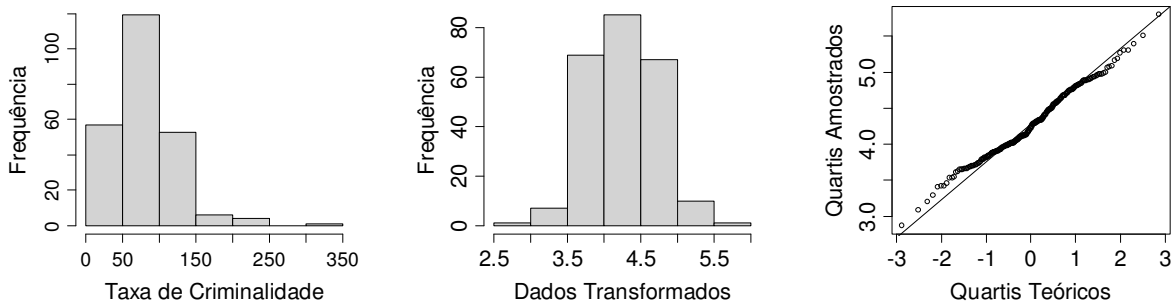
Tabela 4.2 – Modelos Ajustados a cada um dos Casos de Análise.

Caso	Modelo	Base de Dados		Predição	
		Localizações	Tempo	Localizações	Tempo
1	Høst et al. (1995): EQM - Modelo 1	20	12 (1986-1997)	5	12 (1986-1997)
	Høst et al. (1995): *di* - Modelo 2	20	12 (1986-1997)	5	12 (1986-1997)
	Gneiting (2002) – Modelo 3	20	12 (1986-1997)	5	12 (1986-1997)
2	Niu et al. (2003): EQM - Modelo 4	25	9 (1986-1994)	5	3 (1995-1997)
	Niu et al. (2003): *di* - Modelo 5	25	9 (1986-1994)	5	3 (1995-1997)
	Gneiting (2002) – Modelo 6	25	9 (1986-1994)	5	3 (1995-1997)
3	Høst et al. (1995): EQM - Modelo 7	20	9 (1986-1994)	5	9 (1986-1994)
	Høst et al. (1995): *di* - Modelo 8	20	9 (1986-1994)	5	9 (1986-1994)
	Niu et al. (2003): EQM - Modelo 9	25	9 (1986-1994)	5	3 (1995-1997)
	Niu et al. (2003): *di* - Modelo 10	25	9 (1986-1994)	5	3 (1995-1997)
	Gneiting (2002) – Modelo 11	25	9 (1986-1994)	5	3 (1995-1997)

4.3 Análise: Caso 1

Uma análise exploratória dos dados foi realizada inicialmente com o objetivo de se conhecer melhor os dados e auxiliar na implementação dos modelos. Esta análise foi feita usando a base de dados formada pelas 20 regiões administrativas nos anos de 1986 a 1997.

A distribuição da variável taxa de criminalidade é bastante assimétrica como mostra a Figura 4.4 (a). A metodologia de Box-Cox (BOX; COX, 1964) sugeriu aplicar a função logarítmica aos dados para satisfazer a suposição de normalidade. A suposição de normalidade pode ser verificada pela Figura 4.4 (b) e (c).



(a) (b) (c)

Figura 4.4. (a) Histograma dos dados; (b) Histograma dos dados transformados e (c) Q-Q plot do logaritmo da taxa.

A Figura 4.5 mostra o logaritmo da taxa de criminalidade nas 20 regiões administrativas ao longo do tempo. Aparentemente existe uma tendência crescente na taxa de criminalidade.

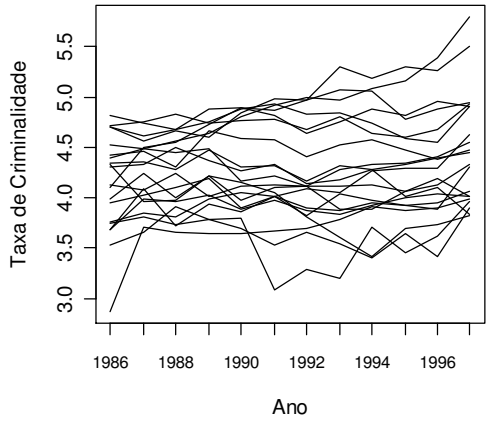


Figura 4.5. Séries temporais das 20 regiões administrativas.

Para cada tempo separadamente aplicou-se as técnicas usuais de estatística espacial e ajustou-se um modelo geoestatístico pelo método de máxima verossimilhança. Esta análise também é um estudo exploratório com o objetivo de entender melhor os dados. O modelo teórico ajustado ao variograma experimental para cada ano foi o cúbico[9] (ver p. 25). A Figura 4.6 apresenta os ajustes dos modelos teóricos aos variogramas amostrais para cada ano.

[9] Em alguns anos o modelo ajustado foi diferente do cúbico. Porém como os valores do AIC, BIC e o logaritmo da verossimilhança eram muito similares entre o modelo inicialmente ajustado e o modelo cúbico, adotamos o último para que fosse possível avaliar a evolução das estimativas dos parâmetros do modelo ao longo do tempo

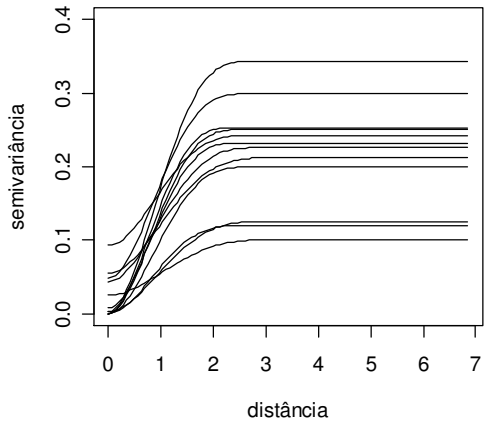


Figura 4.6. Variogramas ajustados para cada tempo separadamente.

Os parâmetros de média, efeito pepita, variância e de escala estimados pelo modelo cúbico para cada um dos tempos separadamente são apresentados na Figura 4.7. A linha horizontal no gráfico corresponde à média das estimativas nos 12 anos investigados. Parece que existe uma tendência crescente nas componentes de média (a) e de variação (c). A análise do parâmetro de escala sugere que a dependência espacial diminui com o passar do tempo (d).

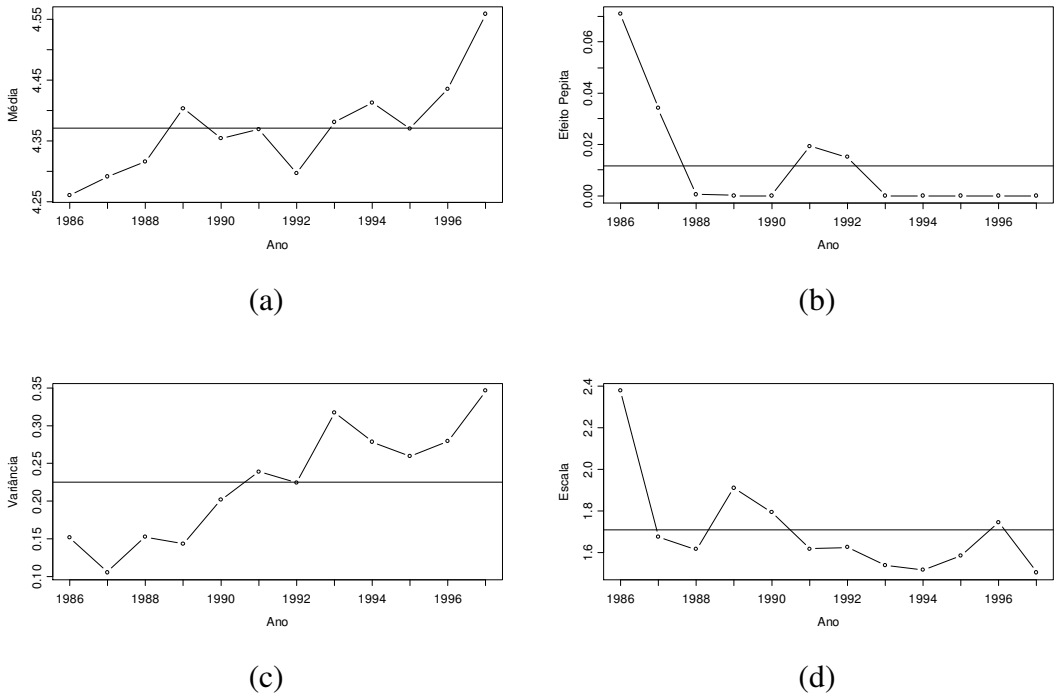


Figura 4.7: Parâmetros estimados para a média (a), efeito pepita (b), variância (c) e escala (d).

O ajuste dos dados pelo modelo proposto por Høst et al. (1995) é apresentado na seção seguinte.

4.3.1 Ajuste pelo Modelo Proposto por Høst et al. (1995)

Nesta seção mostramos o ajuste dos dados pelo modelo proposto por Høst et al. (1995) com os parâmetros estimados pelo método de Kyriakidis e Journel (1999) e com as modificações discutidas no Capítulo 3 (ver seção 3.5.1, p. 53).

O primeiro passo para o ajuste deste modelo aos dados é a identificação do modelo teórico de variograma ajustado a componente *F* e a estimação de seus parâmetros (ver p. 45). O modelo ajustado para esta componente pelo método de máxima verossimilhança foi o cúbico (ver p. 25) e os parâmetros estimados para as componentes de média, efeito pepita, variância e escala são dados por respectivamente: 4,34; 0,0093; 0,17 e 2,88, isto é:

$$\gamma(h) = 0{,}0093 + 0{,}17\begin{cases} 7\left(\frac{h}{2{,}88}\right)^2 - 8{,}75\left(\frac{h}{2{,}88}\right)^3 + 3{,}5\left(\frac{h}{2{,}88}\right)^5 - 0{,}75\left(\frac{h}{2{,}88}\right)^7, & \text{se } h < 2{,}88 \\ 0, & \text{caso contrário} \end{cases} \quad (4.1)$$

O variograma teórico ajustado à componente *F* é apresentado na Figura 4.8.

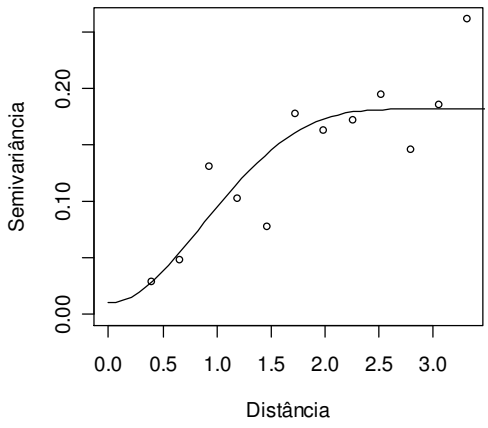


Figura 4.8. Variograma teórico ajustado à componente *F*.

A predição do logaritmo da taxa de criminalidade nos anos de 1986 a 1997 para as cinco regiões administrativas, que foram separadas do banco de dados original para testar o modelo ajustado, foi feita variando-se a quantidade de vizinhos. O número de vizinhos foi calculado de duas maneiras distintas. A primeira forma é baseada no erro quadrático médio, ou seja, o número de vizinhos é identificado pelo menor valor do erro. O erro quadrático médio (EQM) é a média da soma de quadrados da diferença entre o valor observado (ou real) e o valor predito pelo modelo para todas as cinco localizações de validação nos 12 períodos

de tempo. Neste caso utilizamos a mesma quantidade de vizinhos na predição em cada uma das 5 regiões de validação. Esta forma de cálculo não é realista, pois na prática não temos acesso aos valores reais das localizações que estamos querendo predizer em cada tempo. Neste estudo essa forma de cálculo será utilizada apenas para efeito de comparação de modelos e para a avaliação do método da distância *di.*

Os erros correspondentes as predições do logaritmo da taxa de acordo com a quantidade de vizinhos podem ser visualizados na Figura 4.9. Pela análise do EQM o melhor modelo ajustado foi aquele que considera o vizinho mais próximo na predição.

A segunda maneira para calcular a quantidade de vizinhos considera a distância dentre a qual as observações são espacialmente dependentes. Esta distância denotada por *di* é determinada pelo modelo de variograma teórico ajustado à componente *F* e seu valor é igual ao parâmetro de alcance[10] (ou escala) estimado. Neste caso a quantidade de vizinhos não precisa ser a mesma nos cinco locais de predição de validação e o seu valor é determinado pelo número total de localidades que estão a uma distância menor ou igual à *di* do local onde será feita a predição. Esta segunda forma para o cálculo da quantidade de vizinhos é mais realista, visto que não requer o conhecimento dos verdadeiros valores das observações nos locais e tempos de predição.

Para o logaritmo da taxa de criminalidade consideramos que uma distância superior a 2,88 graus implica na independência espacial entre as observações. Esta distância corresponde ao valor do parâmetro de alcance estimado pelo modelo de variograma teórico cúbico ajustado a componente *F* (ver equação (4.1)). O número de vizinhos utilizado na predição do logaritmo da taxa de criminalidade nas regiões de validação Alto do Jequitinhonha, Central, Noroeste, Vale do Rio Grande e Vale do Sapucaí são respectivamente: 10, 14, 2, 5 e 8.

A Figura 4.10 apresenta o erro quadrático médio de acordo com a distância *di*. O ponto em destaque "*" no gráfico corresponde à distância igual a 2,88 graus. Este gráfico não pode ser construído na prática, pois não temos acesso aos valores reais. O objetivo da construção desse gráfico no estudo é de apenas mostrar que *di* igual a 2,88 graus é uma escolha razoável, i.e., as distâncias próximas deste valor resultam em um erro pequeno de predição se comparado com outras distâncias.

[10] Nos casos em que o patamar é atingido assintoticamente o valor de *di* é especificado pelo alcance prático.

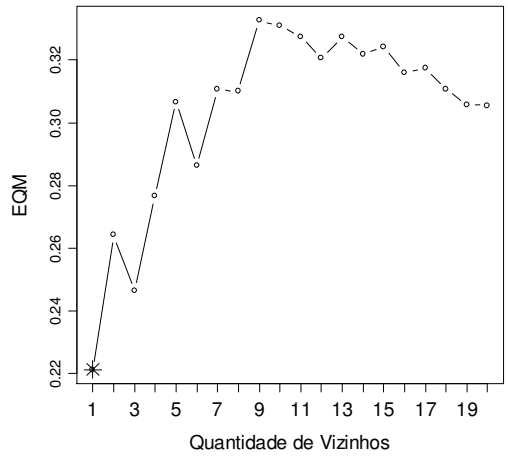


Figura 4.9. EQM de acordo com a quantidade de vizinhos.

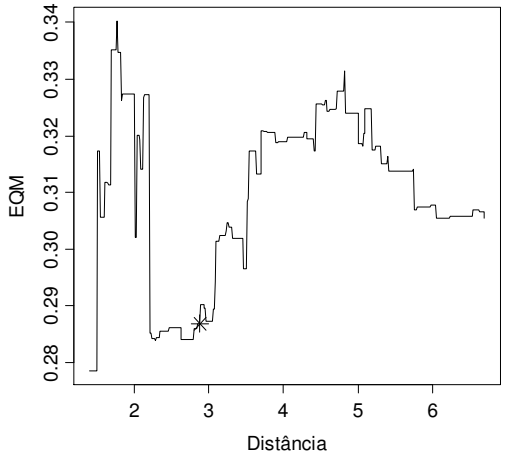


Figura 4.10. EQM de acordo com a distância.

Para facilitar a leitura e a compreensão do texto da dissertação denotaremos o modelo ajustado aos dados baseado no critério do EQM como Modelo 1 e o modelo baseado no critério da distância *di* como Modelo 2. Esta notação vai ser utilizada ao longo deste capítulo.

Para verificar se os Modelos 1 e 2 se ajustaram adequadamente aos dados foi feita à validação cruzada. Este procedimento foi realizado da seguinte maneira: retiramos ponto a ponto do banco de dados que contêm as 20 regiões usadas para o ajuste dos modelos, predizemos o logaritmo da taxa no ponto para todos os tempos amostrados usando o modelo ajustado e calculamos os resíduos. A média dos resíduos calculada pelos Modelos 1 e 2 é próxima de zero e igual à respectivamente: 0,0154 e 0,03. O desvio-padrão dos resíduos é igual a 0,52 e 0,41 para os Modelos 1 e 2 respectivamente. A distribuição dos resíduos para os dois modelos é aproximadamente normal. Concluímos que o ajuste dos Modelos 1 e 2 aos dados do logaritmo da taxa de criminalidade parece adequado.

Os valores preditos pelos Modelos 1 e 2 para as cinco localizações de validação será apresentado na seção 4.3.4, pois estes valores serão comparados com aqueles obtidos pelo ajuste das funções de covariância separável e não-separável da família de Gneiting (2002). A próxima seção apresenta o ajuste por estas funções de covariância.

4.3.2 Ajuste por Funções de Covariância da Família de Gneiting (2002)

A forma da função de covariância espaço–temporal separável ajustada aos dados é dada por:

$$C(h,u) = \frac{\sigma^2}{\left(a|u|^{\alpha}+1\right)^{\delta}} \exp\left(-c\|h\|^{\gamma}\right) + \frac{\sigma^2_{h=0}}{\left(a|u|^{\alpha}+1\right)^{\delta}} \tag{4.2}$$

Esta função é composta pela multiplicação de uma função de covariância Cauchy generalizada (ou gencauchy) (ver p. 26), ajustada ao componente puramente temporal, por uma função de covariância *stable* (ver p. 26), ajustada ao componente puramente espacial. Para este fator é adicionado um efeito de pepita que depende do tempo e corresponde ao produto de uma constante pela função de covariância Cauchy generalizada. Segundo Silva (2006, p. 42) "este modelo é um caso particular da família de funções de covariância espaço–temporais não separáveis propostas por Gneiting, com o valor do parâmetro β igual a zero".

O modelo de covariância não-separável ajustado aos dados é formado pela multiplicação de uma função de covariância Cauchy generalizada, ajustada a parte temporal, por uma função de covariância espaço–temporal não-separável da família de Gneiting. A estas componentes adicionamos um efeito de pepita, puramente espacial, multiplicado por uma função de covariância Cauchy generalizada ajustada à parte temporal. Este modelo de covariância espaço–temporal não-separável pode ser escrito como:

$$C(h,u) = \frac{1}{\left(a|u|^{\alpha}+1\right)^{\delta}} \left(\frac{\sigma^2}{\left(a|u|^{\alpha}+1\right)^{\beta}} \exp\left(-c \left[\frac{\|h\|}{\left(a|u|^{\alpha}+1\right)^{\beta/2}} \right]^{\gamma} \right) \right) + \frac{\sigma^2_{h=0}}{\left(a|u|^{\alpha}+1\right)^{\delta}} \tag{4.3}$$

Quando o valor de β é igual à zero, o modelo não separável (4.3) se reduz ao modelo separável (4.2). O parâmetro β assume valores no intervalo $[0,1]$. Os valores de β próximos de 1 indicam forte interação espaço–temporal. Para valores de β próximos de zero, temos a indicação de que a dependência espaço–temporal é fraca, ou seja, os dois processos espaço e tempo atuam de forma independente.

A função de covariância espaço-temporal não-separável da família de Gneiting que compõe a fórmula (4.3) é na verdade a função de covariância *nsst* (*Non-Separable Space-Time model*). De acordo com a documentação do pacote *RandomFields* do *software* R esta função pode ser escrita como:

$$C(h,u) = \psi(u)^{-f} \phi\left(\frac{h}{\psi(u)} \right) \tag{4.4}$$

onde f deve ser maior que a dimensão espacial do processo. A função ϕ utilizada neste capítulo é a função *stable*, e a função $\psi^2(u)$ é igual a $\left(u^c+1\right)^d$, onde $c \in (0,2]$ e $d \in [0,1]$.

As funções de covariância espaço–temporais mostradas em (4.2) e (4.3) foram ajustadas aos dados. Estas funções foram escolhidas pela acessibilidade de implementação, visto que Silva (2006) utiliza estas covariâncias para ajustar dois bancos de dados distintos obtendo resultados satisfatórios. Um dos objetivos do trabalho de pesquisa do autor foi estudar o pacote *RandomFields* do *software* R e verificar como é feita a implementação destas funções da família de Gneiting.

O modelo separável da forma apresentada em (4.2) ajustado aos dados do logaritmo da taxa de criminalidade pelo método de máxima verossimilhança é dado em (4.5):

$$C(h,u) = \frac{2{,}26}{\left(0{,}0698|u|^{1{,}98}+1\right)^{1{,}34}} \exp\left(-0{,}0146\|h\|^{0{,}65}\right) + \frac{0{,}0140}{\left(0{,}0698|u|^{1{,}98}+1\right)^{1{,}34}} \tag{4.5}$$

O valor do parâmetro β estimado pelo modelo não-separável da forma (4.3) ajustado aos dados do logaritmo da taxa é igual a zero. Dessa forma, o modelo não-separável se reduz ao modelo separável dado em (4.5). Sendo assim as predições do logaritmo da taxa nas cinco regiões administrativas de validação foram calculadas usando o modelo de covariância espaço–temporal separável dado em (4.5), pois o valor de β sugere a independência entre as estruturas temporal e espacial. O ajuste do modelo (4.5) aos dados é denotado por Modelo 3.

De acordo com Silva (2006) pode ocorrer dificuldades nos algoritmos numéricos na identificação do parâmetro β. Uma forma de contornar esse problema é considerar duas estratégias distintas: a verossimilhança condicional e a perfilhada, onde ambas "isolam" a estimação de β.

Seja (θ,β) o vetor de parâmetros em que θ é o vetor que contêm todos os parâmetros do modelo exceto β. Na verossimilhança condicional tomamos uma seqüência de valores para β e obtemos a função de verossimilhança fixando os parâmetros θ em suas estimativas de verossimilhança, i.e., $L(\beta,\theta) = \hat{\theta}$. Na verossimilhança perfilhada obtemos para cada valor de β a verossimilhança maximizada em relação aos demais parâmetros em θ, ou seja, $L(\beta,\theta) = \hat{\theta}_\beta$.

A Figura 4.11 mostra a verossimilhança condicional e a verossimilhança perfilhada utilizando o logaritmo dos dados de criminalidade.

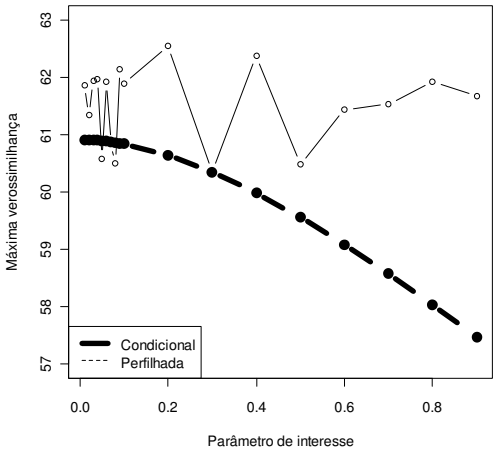


Figura 4.11. Máxima verossimilhança condicional e perfilhada.

A verossimilhança condicional apresenta o maior valor da função em $\beta = 0$ indicando um modelo de covariância separável. O máximo da verossimilhança perfilhada ocorre quando $\beta = 0{,}2$, porém a diferença entre este valor e aquele quando $\beta = 0$ é pequena. O estudo das verossimilhanças confirma a hipótese de instabilidade do parâmetro β e conseqüentemente a dificuldade na sua estimação.

A próxima seção apresenta os resultados das predições utilizando os três modelos. Os Modelos 1 e 2 apresentados na seção 4.3.1 se referem à proposta modificada de Høst et al. (1995) utilizando respectivamente, o critério do erro quadrático médio (EQM) e o critério da distância *di* para a escolha do número de vizinhos para predição, e o Modelo 3 é concernente à função de covariância separável pertencente à família de Gneiting dada em (4.5).

4.3.3 Comparação dos Modelos Ajustados: Modelos 1, 2 e 3

A Tabela 4.3 apresenta o erro quadrático médio (EQM) e a média dos resíduos (RES) de acordo com o modelo ajustado para o logaritmo da taxa. Estes erros foram calculados usando a predição do logaritmo da taxa nas 5 regiões do banco de validação nos 12 instantes de tempo (1986 a 1997). O Modelo 1 se refere a proposta de Høst et al. (1995) que utiliza o vizinho mais próximo para a predição, o Modelo 2 é aquele que utiliza a distância *di* igual a 2,88 graus para o cálculo do número de vizinhos e o Modelo 3 se refere a função de covariância separável da família de Gneiting (2002) mostrada em (4.5).

Tabela 4.3 – Comparação dos modelos 1, 2 e 3.

Modelo	EQM	RES
Høst et al. (1995): EQM - Modelo 1	0,2214	0,1666
Høst et al. (1995): *di* - Modelo 2	0,2868	0,2254
Gneiting (2002): Separável - Modelo 3	0,3524	0,3037

O Modelo 1 foi o que apresentou o melhor ajuste, isto é, obtemos o menor erro quadrático médio e a menor média residual se comparado com os outros modelos. O Modelo 3 baseado na função de covariância não-separável da família de Gneiting foi o que apresentou o pior ajuste.

A Figura 4.12 apresenta o erro quadrático médio (EQM) por região administrativa para os três modelos ajustados. O erro é calculado utilizando as predições do log da taxa de criminalidade nos anos de 1986 a 1997 para cada uma das regiões de validação. A região Central é a que apresenta o valor do erro mais alto se comparado com as outras regiões. Esta região reúne cidades populosas e com muita variação nas taxas de criminalidade com o passar dos anos, por exemplo, a capital do estado de Minas Gerais, Belo Horizonte (ver Tabela 4.1).

O erro quadrático médio do logaritmo da taxa por ano de acordo com o modelo ajustado é mostrado na Figura 4.13. Na média o comportamento dos erros, calculados pela predição do logaritmo da taxa de criminalidade nas cinco localizações de validação em cada um dos anos, é semelhante para os três modelos ao longo do tempo. Observamos um aumento no erro de predição para os três modelos a partir do ano de 1990, e um decréscimo entre os anos de 1986 e 1989. O valor alto do erro para o ano de 1986 utilizando o Modelo 2 corresponde a predição para a região administrativa Central. Os erros obtidos pelo Modelo 3 são relativamente maiores se comparado com os outros modelos no período analisado.

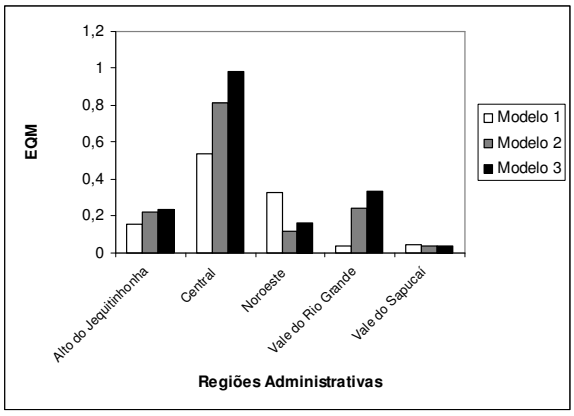


Figura 4.12. Erro quadrático médio por região administrativa de acordo com o modelo ajustado.

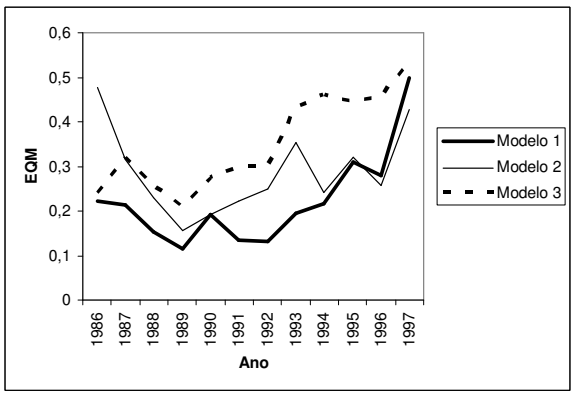


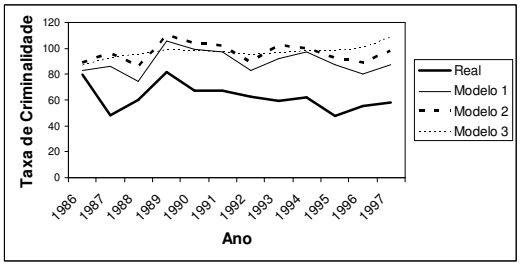
Figura 4.13. Erro quadrático médio por ano de acordo com o modelo ajustado.

Os Quadros de 4.1 a 4.5 apresentam os resultados das predições para as cinco regiões administrativas de validação no período entre 1986 e 1997 considerando os três modelos descritos anteriormente para o ajuste. Os valores apresentados nas tabelas e nos gráficos estão na escala original da variável taxa de criminalidade, ou seja, aplicou-se a transformação inversa (exponencial) aos dados preditos.

Observamos que os valores preditos pelos três modelos foram superestimados para a região Alto do Jequitinhonha. Apesar disto, os Modelos 1 e 2 conseguiram acompanhar o comportamento da série de valores reais ao longo dos anos. O ajuste dos modelos na região Central não foi bom, pois os valores foram subestimados. Na região Noroeste as predições do período inicial e final utilizando os Modelos 1 e 2 conseguem acompanhar o comportamento da série real, apesar de serem subestimados. Parece que os modelos não conseguem captar essa grande variação na taxa de criminalidade entre os anos de 1991 e 1993. Os três modelos ajustados para a região do Vale do Rio Grande apresentaram uma tendência crescente da taxa de criminalidade, assim como a série de valores observados. As predições feitas pelo Modelo 1 foram as que apresentaram os melhores resultados. As predições feitas no Vale do Sapucaí foram consideradas satisfatórias, visto que todos os modelos conseguiram reproduzir a tendência crescente apresentada na série real.

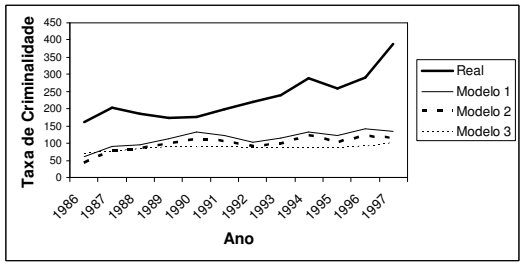
Quadro 4.1. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa Alto do Jequitinhonha.

Alto do Jequitinhonha				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1986	79,26	82,75	88,7	86,63
1987	47,95	85,96	95,58	92,87
1988	60,02	74,61	86,34	95,33
1989	81,24	105,9	109,91	98,36
1990	67,45	98,81	103,37	98,05
1991	67,09	96,97	101,81	96,87
1992	62,68	82,71	89,52	95,29
1993	59,26	92,16	102,67	96,78
1994	61,73	96,85	99,62	98,09
1995	47,56	87,41	91,92	97,66
1996	55,22	80,35	88,44	100,53
1997	58	87,19	97,9	108,35
Erro médio		-27,02	-34,03	-34,78
$\sqrt{\text{EQM}}$		28,78	35,44	36,89



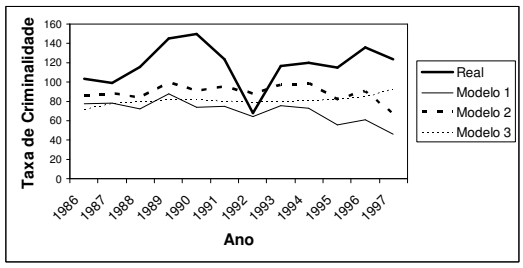
Quadro 4.2. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa Central.

Central				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1986	160,89	60,38	41,75	68,61
1987	202,61	90,16	78,05	77,01
1988	184,68	94,48	83,24	82,17
1989	174,34	112,38	96,81	87,19
1990	176,13	132,98	112,09	88,08
1991	197,44	123,12	106,03	86,99
1992	219,12	103,19	90,47	84,84
1993	240,85	115,48	96,93	85,32
1994	287,75	132,24	122,2	86,19
1995	258,94	122,84	103,07	86,34
1996	290,1	143,06	119,74	90,24
1997	389,46	135,17	114,58	99,61
Erro médio		118,07	134,78	146,64
$\sqrt{\text{EQM}}$		129,18	144,92	157,83



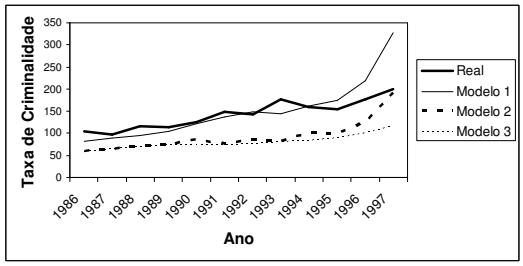
Quadro 4.3. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa Noroeste.

Noroeste				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1986	103,16	76,98	86,18	71,38
1987	99,3	78,04	88	77,07
1988	115,9	72,23	83,41	79,42
1989	145,41	87,77	99,53	81,9
1990	149,4	74,29	90,13	81,38
1991	123,49	75,12	94,41	79,99
1992	67,66	64,67	87,76	78,48
1993	116,85	75,49	96,22	79,69
1994	119,77	72,79	98,06	81,13
1995	114,65	55,65	81,75	81,36
1996	135,89	60,54	90,73	84,48
1997	123,57	46,35	67,88	91,82
Erro médio		47,93	29,25	37,25
$\sqrt{\text{EQM}}$		52,75	35,89	41,97



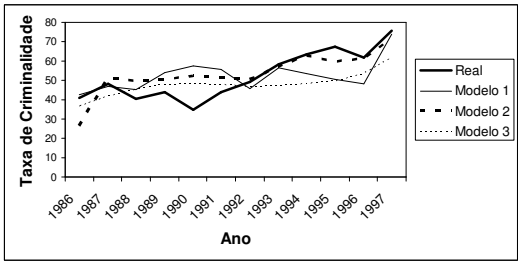
Quadro 4.4. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa do Vale do Rio Grande.

Vale do Rio Grande				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1986	103,72	81,5	58,69	58,53
1987	96,15	88,54	63,87	65,36
1988	115,93	95,65	71,5	69,41
1989	115,08	103,98	75,06	73,44
1990	124,86	121,37	84,9	74,56
1991	149,01	137,68	76,08	74,76
1992	141,83	147,65	84,72	75,23
1993	176,9	144,16	81,26	79,13
1994	160,47	162,51	101,73	84,54
1995	154,02	175,26	99,29	89,95
1996	176,07	218,85	124,79	99,91
1997	200,17	327,86	191,46	116,56
Erro médio		-7,57	50,07	62,74
$\sqrt{\text{EQM}}$		41,76	54,10	65,57



Quadro 4.5. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa do Vale do Sapucaí.

Vale do Sapucaí				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1986	41,05	42,67	26,46	36,69
1987	48,17	46,75	51,18	41,79
1988	40,29	45,15	49,7	45,05
1989	43,85	53,94	50,47	47,97
1990	34,77	57,36	52,3	48,42
1991	44,1	55,64	51,29	47,65
1992	48,92	45,57	50,47	46,53
1993	58,19	56,5	57	47,13
1994	63,39	53,48	62,66	48,37
1995	67,18	50,53	59,49	49,66
1996	61,67	48,47	61,26	53,64
1997	75,6	74,12	71,68	61,71
Erro médio		-0,25	-1,40	4,38
$\sqrt{\text{EQM}}$		10,57	8,13	10,07



Observamos que as predições são melhores (erros menores) naqueles locais onde os vizinhos têm comportamento semelhante com a localidade de predição (ver Figura 4.3, p. 62). Os resultados mais favoráveis correspondem às regiões administrativas do Vale do Rio Grande e do Vale do Sapucaí. A região Central é a que apresenta o pior ajuste. Aparentemente os modelos ajustados não conseguem captar grandes variações na série de dados.

A Tabela 4.4 resume as principais estatísticas descritivas dos erros globais associados aos três modelos. Os erros são calculados pela diferença entre os valores observados e preditos para todos os anos e as cinco regiões de validação consideradas na predição. Estes

erros também foram calculados na escala original aplicando-se a transformação inversa aos dados preditos.

Tabela 4.4 – Análise descritiva dos erros.

Método	Média	Média absoluta	Desvio padrão	Mínimo	Máximo
Høst et al. (1995): EQM – Modelo 1	26,23	45,38	61,69	-127,69	254,29
Høst et al. (1995): *di* – Modelo 2	35,73	51,53	64,02	-47,63	274,88
Gneiting (2002): Separável – Modelo 3	43,25	58,39	68,52	-50,35	289,85

Os melhores ajustes correspondem aos Modelos 1 e 2, i.e, estes modelos resultam em menores erros em média se comparado com o Modelo 3.

A seção seguinte trata de modelos ajustados aos dados do log da taxa cujo objetivo é a predição temporal da característica de interesse em localizações observadas na amostra.

4.4 Análise: Caso 2

No Caso 2 de análise o objetivo também é predizer a taxa de criminalidade nas regiões Alto do Jequitinhonha, Central, Noroeste, Vale do Rio Grande e Vale do Sapucaí, porém esta predição foi feita somente para os três últimos anos. Escolhe-se somente três anos para testar o ajuste do modelo, pois a série temporal de observações é pequena.

O banco de dados utilizado considera as 25 localizações e para cada localização tem-se a informação da taxa de criminalidade no período de 1986 a 1994. Uma análise descritiva dos dados foi realizada utilizando esta base de dados e os resultados foram muito similares com aqueles apresentados no início da seção 4.3. Aplicou-se também a transformação logarítmica aos dados e os dados transformados têm uma distribuição aproximadamente normal.

A predição destas observações nos anos de 1995 a 1997 foi calculada pelo modelo baseado na proposta de Niu et al. (2003) definido na equação (3.44) (ver p. 54) e ajustaram-se também funções de covariância espaço-temporal da família de Gneiting (2002). Os resultados do ajuste dos modelos são apresentados nas seções subseqüentes.

4.4.1 Ajuste pelo Modelo Baseado na Proposta de Niu et al. (2003)

Nesta seção mostramos o ajuste dos dados pelo modelo baseado na proposta de Niu et al. (2003) e dado pela equação (3.44) (ver seção 3.5, p. 54). A quantidade de vizinhos do

modelo definido na equação (3.44) foi determinada de maneira similar a aquela apresentada no ajuste do modelo de Høst et al. (1995) (ver p. 66). Utilizou-se também o critério do erro quadrático médio (EQM) e o da distância *di*, porém o EQM é calculado usando as predições do logaritmo da taxa de criminalidade nas cinco localidades de validação em um único instante de tempo (no primeiro ano de predição) e a distância *di* é definida pelo variograma teórico ajustado aos dados no tempo imediatamente anterior ao primeiro ano de predição. Como ressaltado na análise do Caso 1 o cálculo do número de vizinhos pelo EQM não é comum na prática, pois para obter o seu valor usamos os valores reais. Abordamos esta forma de cálculo na dissertação para efeito de comparação.

O gráfico do EQM para o log da taxa de acordo com a quantidade de vizinhos pode ser visualizado na Figura 4.14. O ponto em destaque "*" representa o menor valor do EQM e corresponde a uma vizinhança formada por 6 localizações. No cálculo do EQM usamos as predições do log da taxa nas cinco localidades do banco de validação no ano de 1995.

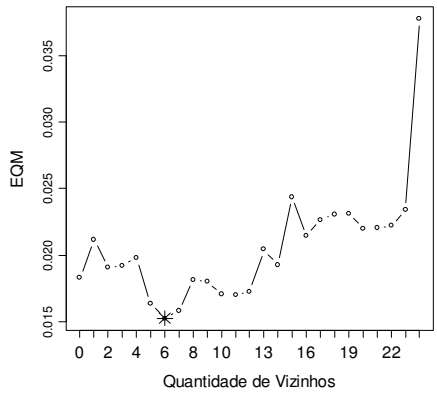


Figura 4.14. EQM de acordo com o número de vizinhos no ano de 1995.

O modelo ajustado segundo as idéias de Niu et al. (2003) aos dados do logaritmo da taxa de criminalidade chamada aqui de *Z,* baseado no critério do EQM para o cálculo do número de vizinhos é dado em (4.6).

$$\begin{aligned} Z(s,t) = \; & 0{,}3671 + 0{,}9032 \times Z(s,t-1) + 0{,}0949 \times Z(v_1,t-1) - 0{,}0349 \times Z(v_2,t-1) \\ & - 0{,}0111 \times Z(v_3,t-1) + 0{,}0160 \times Z(v_4,t-1) - 0{,}0371 \times Z(v_5,t-1) \\ & - 0{,}0117 \times Z(v_6,t-1), \quad t = 1995, 1996, 1997 \end{aligned} \tag{4.6}$$

A equação (4.6) diz que a predição temporal nos anos de 1995, 1996 e 1997 para as cinco localizações desconsideradas na análise para validação, Alto do Jequitinhonha, Central, Noroeste, Vale do Rio Grande e Vale do Sapucaí, depende das informações dos 6 vizinhos

mais próximos (no espaço) e do próprio local de predição no tempo anterior ao da previsão, adicionada de uma constante igual a 0,3671. O peso da informação da característica de interesse no local de predição no tempo anterior ao da previsão é alto e igual a 0,9032.

Os resultados da predição para o logaritmo da taxa de criminalidade aplicando o modelo (4.6) são resumidos na Tabela 4.5. Essa tabela mostra o EQM e a média dos resíduos (RES) de acordo com o ano de predição, e a média da soma de quadrados do erro (MSQE) definida na equação (3.46) (ver p. 55). Ressalta-se que o EQM e o RES são calculados usando as predições do logaritmo da taxa nas cinco localizações de validação para cada um dos anos.

Tabela 4.5 – Resultados da predição pelo ajuste do modelo dado em (4.6).

Tempo	EQM	RES	MSQE
1995	0,0153	-0,0800	0,0260
1996	0,0167	0,0115	----
1997	0,0327	0,1207	----

A segunda forma para determinar o número de vizinhos para predizer as observações nas cinco localidades de validação nos três últimos anos utiliza o critério da distância *di* para o cálculo do número de vizinhos, i.e., deve-se ajustar um modelo de variograma teórico aos dados no ano de 1994 (ano anterior ao primeiro de predição) e especificar a distância *di* que é igual ao parâmetro de escala (ou alcance) estimado. A seguir calcula-se o número de localizações que estão a uma distância menor ou igual a *di* em cada ponto que se deseja fazer a predição temporal, e toma-se a média como sendo a quantidade de vizinhos que deve ser usada nas predições.

O modelo de variograma teórico ajustado aos dados no ano de 1994 pelo método de máxima verossimilhança foi o circular (ver p. 25) e os parâmetros estimados para as componentes de média, efeito pepita, variância e escala, são respectivamente: 4,41; 0; 0,28 e 1,52, ou seja:

$$\gamma(h) = \begin{cases} 0{,}28(\Gamma(h)) & , h > 0 \\ 0 & , h \leq 0 \end{cases} \tag{4.7}$$

onde $\Gamma(h) = \dfrac{2\left[\left(\theta\sqrt{1-\theta^2}\right) + sen^{-1}\sqrt{\theta}\right]}{\pi}$ e $\theta = \min\left(\dfrac{h}{1{,}52}, 1\right)$.

O modelo de variograma teórico ajustado aos dados no ano de 1994 é mostrado na Figura 4.15.

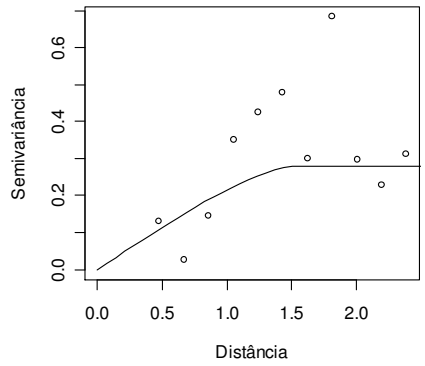


Figura 4.15. Variograma teórico ajustado aos dados no ano de 1994.

Neste estudo consideramos que no ano de 1994 não existe dependência espacial entre as observações separadas por uma distância superior a 1,52 graus. Esta distância corresponde ao parâmetro de alcance estimado pelo variograma circular ajustado aos dados e implica em um número de vizinhos igual a 3 para predizer as observações nos anos de 1995, 1996 e 1997 nas cinco localidades de validação. O modelo estimado usando o critério da distância *di* para quantificar os vizinhos pode ser escrito como em (4.8):

$$\begin{aligned} Z(s,t) = \quad & 0{,}3392 + 0{,}8938 \times Z(s,t-1) + 0{,}0917 \times Z(v_1,t-1) - 0{,}0550 \times Z(v_2,t-1) \\ & - 0{,}0052 \times Z(v_3,t-1), \quad t = 1995,1996,1997 \end{aligned} \tag{4.8}$$

A Figura 4.16 mostra o EQM de acordo com a distância e este erro é baseado nas predições das cinco localidades de validação no ano de 1995. O ponto em destaque "*" corresponde ao erro associado à distância igual a 1,52 graus. Este gráfico meramente ilustrativo objetiva verificar a adequação da escolha de *di.* Em situações reais este gráfico não pode ser construído, pois não temos acesso aos valores verdadeiros da característica de interesse. Observamos que a distância igual a 1,52 graus é razoável. Valores de distância pertencentes ao intervalo [2;3] graus fornecem erros menores de predição.

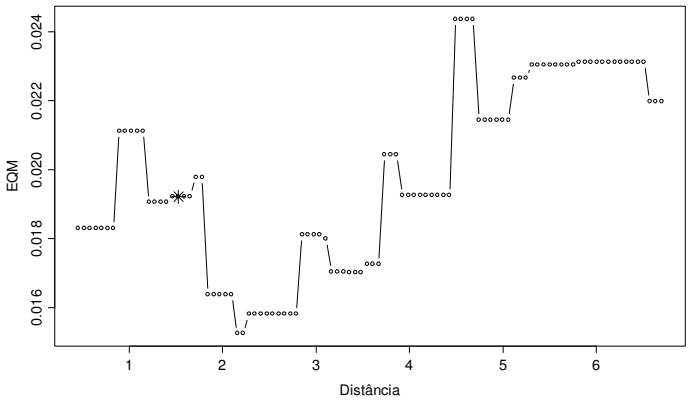


Figura 4.16. Erro quadrático médio de acordo com a distância para o ano de 1995.

A Tabela 4.6 apresenta os resultados das predições do logaritmo da taxa de criminalidade nas cinco localidades de validação calculadas pelo modelo dado em (4.8) de acordo com o ano. Essa tabela mostra o EQM e a média dos resíduos (RES) de acordo com o ano de predição, e a média da soma de quadrados do erro (MSQE) definida na equação (3.46) (ver p. 55).

Tabela 4.6 – Resultados da predição pelo ajuste do modelo dado em (4.8).

Tempo	EQM	RES	MSQE
1995	0,0192	-0,0718	0,0264
1996	0,0254	0,0287	----
1997	0,0590	0,1473	----

Os modelos baseados no critério do EQM e na distância *di* para o cálculo do número de vizinhos e que fazem a predição temporal nas localizações amostradas são denotadas neste trabalho por, respectivamente: Modelo 3 e Modelo 4.

Observamos pelas Tabelas 4.5 e 4.6 que os erros de predição obtidos pelo ajuste do Modelo 3 são menores se comparados com o ajuste pelo Modelo 4, porém o número de parâmetros do Modelo 4 é menor que do Modelo 3. O Modelo 4 utiliza os 3 vizinhos mais próximos na predição, já o Modelo 3 necessita de 6 vizinhos.

A seção seguinte apresenta o ajuste por funções de covariância da família de Gneiting (2002). Os valores preditos da taxa de criminalidade na escala original pelo ajuste dos Modelos 3 e 4, e por estas funções de covariância são mostradas na seção 4.4.3 onde é feita a comparação destes modelos ajustados.

4.4.2 Ajuste por Funções de Covariância da Família de Gneiting (2002)

As funções de covariância espaço–temporal separável dada pela equação (4.2) e não-separável dada pela equação (4.3) (ver p. 69) da família de Gneiting estimadas pelo método de máxima verossimilhança utilizando os dados do período de 1986 a 1994 e as 25 localizações são dadas respectivamente por (4.9) e (4.10):

$$C(h,u)=\frac{2{,}46}{\left(0{,}0348|u|^{1{,}89}+1\right)^{1{,}90}}\exp\left(-0{,}0146\|h\|^{0{,}56}\right)+\frac{0{,}0173}{\left(0{,}0348|u|^{1{,}89}+1\right)^{1{,}90}} \tag{4.9}$$

$$\begin{aligned} C(h,u)=&\left(\frac{2{,}46}{\left(0{,}0348|u|^{1{,}89}+1\right)^{0{,}0782}}\exp\left(-0{,}0146\left[\frac{\|h\|}{\left(0{,}0348|u|^{1{,}89}+1\right)^{0{,}0391}}\right]^{0{,}56}\right)\right) \\ &\times\frac{1}{\left(0{,}0348|u|^{1{,}89}+1\right)^{1{,}90}}+\frac{0{,}0173}{\left(0{,}0348|u|^{1{,}89}+1\right)^{1{,}90}} \end{aligned} \tag{4.10}$$

O parâmetro β estimado em (4.9) é igual a 0,0782 indicando que existe uma fraca interação entre os processos espacial e temporal. Aparentemente os dois processos não atuam simultaneamente e não devem ser considerados como processos dependentes.

O gráfico da verossimilhança condicional e perfilhada é mostrado na Figura 4.17. Observamos que a verossimilhança condicional fornece conclusões similares com aquelas obtidas anteriormente, ou seja, temos indicação que o modelo separável deve ser ajustado aos dados. O máximo da função de verossimilhança perfilhada ocorre quando $\beta=0{,}4$. Pelo estudo das verossimilhanças concluímos que o parâmetro β é muito instável e difícil de ser estimado.

A predição temporal será calculada usando o modelo separável dado em (4.9) e o ajuste denotado por Modelo 6.

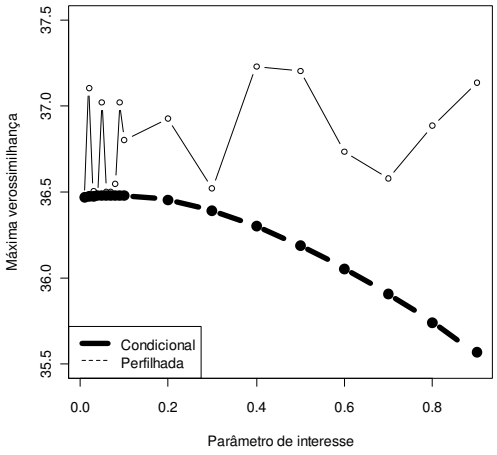


Figura 4.17. Máxima verossimilhança condicional e perfilhada.

A comparação dos ajustes dos dados pelos Modelos 4, 5 e 6 é mostrada na próxima seção.

4.4.3 Comparação dos Modelos Ajustados: Modelos 4, 5 e 6

A Tabela 4.7 compara os ajustes dos dados pelos Modelos 4, 5 e 6, sendo os dois primeiros modelos baseados na proposta de Niu et al. (2003) e que utilizam respectivamente o critério do EQM e a distância *di* para o cálculo do número de vizinhos e o último, o Modelo 6, se refere ao ajuste pela função de covariância separável da família de Gneiting (2002) dada pela equação (4.9). A Tabela 4.7 apresenta o erro quadrático médio (EQM) e a média dos resíduos (RES) calculados usando as predições do logaritmo da taxa nas cinco localizações de validação nos anos de 1995, 1996 e 1997.

Tabela 4.7 – Comparação dos modelos 4, 5 e 6.

Modelo	EQM	RES
Niu et al. (2003): EQM - Modelo 4	0,0216	0,0174
Niu et al. (2003): *di* - Modelo 5	0,0345	0,0347
Gneiting (2002): Separável - Modelo 6	0,0695	0,0748

Os Modelos 4 e 5 apresentaram um melhor ajuste se comparados com o Modelo 6, pois os erros das predições são menores.

A Figura 4.18 apresenta o erro quadrático médio de acordo com a região administrativa para os três tempos (1995, 1996 e 1997). Os erros de predição pelo ajuste do

Modelo 6 são maiores para todas as regiões, exceto para o Noroeste, se comparados com os Modelos 4 e 5. A região Central é a que apresenta os maiores erros de predição e os menores erros correspondem às regiões do Vale do Rio Grande e do Vale do Sapucaí.

A Figura 4.19 mostra o EQM por ano para os três modelos. O comportamento dos erros nos três anos é semelhante entre os modelos. A pior predição é no ano de 1997 como esperávamos, pois esta predição é baseada em predições de dois anos anteriores.

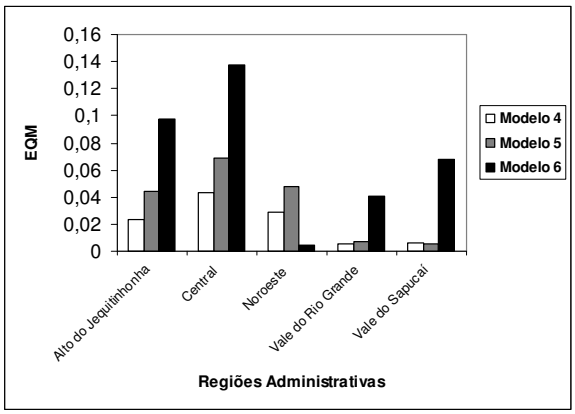


Figura 4.18. Erro quadrático médio por região administrativa de acordo com o modelo ajustado.

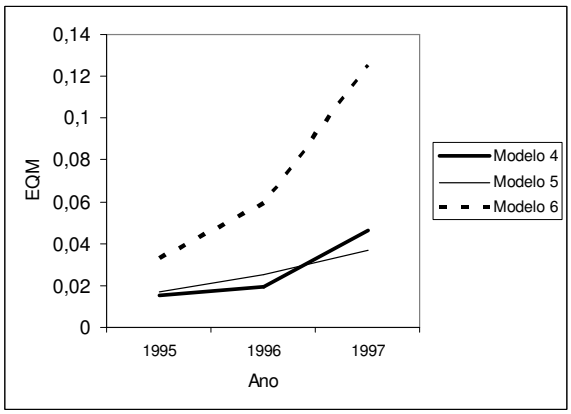


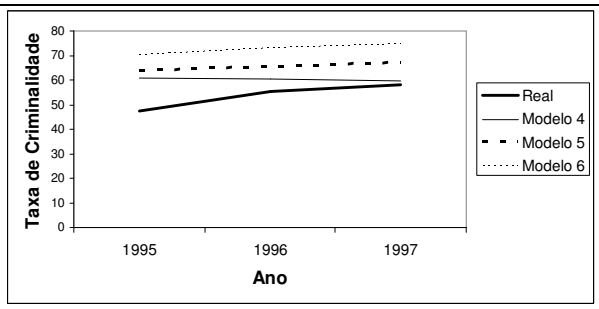
Figura 4.19. Erro quadrático médio por ano de acordo com o modelo ajustado.

Os Quadros de 4.6 a 4.10 apresentam os valores observados e os valores preditos pelos Modelos 4, 5 e 6 da taxa de criminalidade em cada uma das cinco regiões de validação nos anos de 1995, 1996 e 1997. Aplicou-se a função inversa (função exponencial) aos valores preditos pelos modelos.

As melhores predições correspondem às regiões do Vale do Sapucaí e do Vale do Rio Grande. Os Modelos 4 e 5 foram os que apresentaram os melhores ajustes, exceto para a região Noroeste.

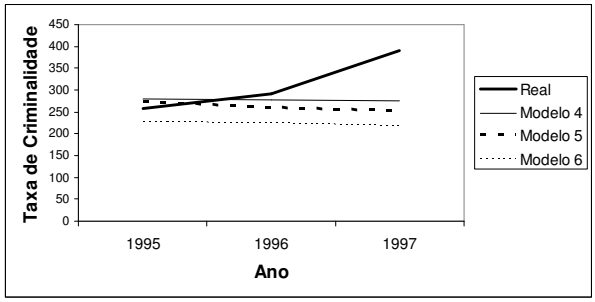
Quadro 4.6. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa Alto do Jequitinhonha.

Alto do Jequitinhonha				
Ano	Real	Modelo 4	Modelo 5	Modelo 6
1995	47,56	60,90	63,58	70,34
1996	55,22	60,22	65,31	72,78
1997	58,00	59,71	66,98	74,70
Erro médio		-6,68	-11,70	-19,01
$\sqrt{\text{EQM}}$		8,29	12,10	19,20



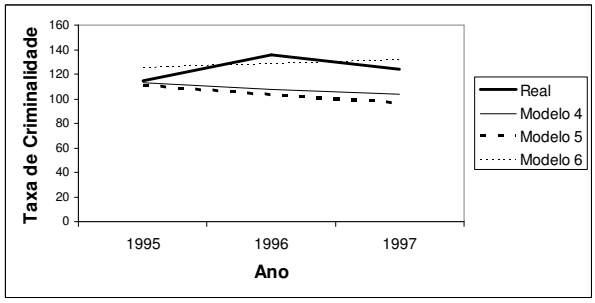
Quadro 4.7. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa Central.

Central				
Ano	Real	Modelo 4	Modelo 5	Modelo 6
1995	258,94	280,86	272,65	227,41
1996	290,10	276,86	260,79	225,16
1997	389,46	274,84	251,17	219,25
Erro médio		35,31	51,30	88,89
$\sqrt{EQM}$		67,81	82,00	106,74



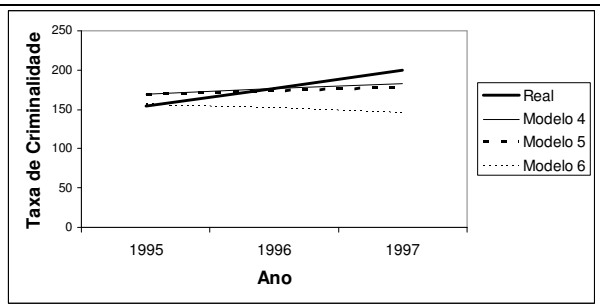
Quadro 4.8. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa Noroeste.

Noroeste				
Ano	Real	Modelo 4	Modelo 5	Modelo 6
1995	114,65	112,87	110,31	124,78
1996	135,89	107,51	102,61	128,97
1997	123,57	103,28	96,18	131,95
Erro médio		16,82	21,67	-3,86
$\sqrt{EQM}$		20,17	25,01	8,58



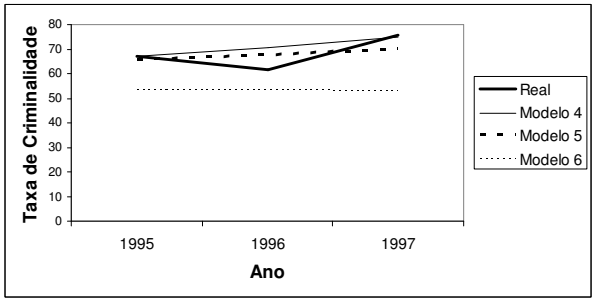
Quadro 4.9. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa do Vale do Rio Grande.

Vale do Rio Grande				
Ano	Real	Modelo 4	Modelo 5	Modelo 6
1995	154,02	168,68	167,39	155,75
1996	176,07	175,94	173,26	152,38
1997	200,17	182,29	178,24	145,92
Erro médio		1,12	3,79	25,40
$\sqrt{EQM}$		13,35	14,92	34,19



Quadro 4.10. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa do Vale do Sapucaí.

Vale do Sapucaí				
Ano	Real	Modelo 4	Modelo 5	Modelo 6
1995	67,18	66,92	65,37	53,36
1996	61,67	70,75	67,64	53,46
1997	75,60	74,79	70,11	52,75
Erro médio		-2,67	0,45	14,96
$\sqrt{\text{EQM}}$		5,26	4,80	16,13



A Tabela 4.8 apresenta as principais medidas estatísticas para resumir as informações dos erros de predição associadas aos três modelos: 4, 5 e 6. Estes erros globais foram calculados utilizando os dados na escala original. O erro é a diferença entre o valor real e o valor predito pelo modelo para os anos de 1995, 1996 e 1997 nas 5 localidades de validação.

Tabela 4.8 – Análise descritiva dos erros.

Método	Média	Média absoluta	Desvio padrão	Mínimo	Máximo
Niu et al. (2003): EQM – Modelo 4	8,78	17,54	32,38	-21,92	114,62
Niu et al. (2003): *di* – Modelo 5	13,1	22,19	38,41	-16,02	138,29
Gneiting (2002): Separável – Modelo 6	21,28	31,58	48,56	-22,78	170,21

Observamos que os Modelos 4 e 5 foram os que apresentaram os melhores ajustes para a taxa de criminalidade em termos dos erros globais médios de predição.

Na próxima seção as metodologias apresentadas nos Casos 1 e 2 de análise (seções 4.3 e 4.4) são combinadas com o intuito de predizer a taxa de criminalidade em localizações e tempos não observados na amostra.

4.5 Análise: Caso 3

No Caso 3 de análise o objetivo é predizer a taxa de criminalidade em localizações e tempos não observados na amostra. As regiões consideradas desconhecidas para validação são as mesmas apresentadas anteriormente nos Casos 1 e 2 de análise: Alto do Jequitinhonha, Central, Noroeste, Vale do Rio Grande e Vale do Sapucaí. Os instantes de tempo de predição são os anos de 1995, 1996 e 1997.

O banco de dados utilizado considera 20 localizações e tem-se a informação da taxa de criminalidade nos anos de 1986 a 1994. A análise descritiva dos dados forneceu resultados

semelhantes com aqueles apresentados na seção 4.3. Aplicou-se a transformação logarítmica aos dados e os dados transformados têm uma distribuição aproximadamente normal.

As metodologias de geoestatística e de séries temporais foram combinadas em duas etapas para predizer estas observações consideradas desconhecidas. Na etapa 1 o modelo de geoestatística de Høst et al. (1995) (ver seção 3.5.1, p. 53) foi ajustado aos dados com o objetivo de se predizer a taxa nas 5 localidades de validação nos anos de 1986 a 1994. Na etapa 2 o banco de dados foi atualizado com as predições calculadas na etapa 1 para as 5 localidades e ajustou-se o modelo baseado nas idéias de Niu et al. (2003) (ver seção 3.5.2, p. 54) para predizer a característica de interesse nos anos de 1995, 1996 e 1997 para estas mesmas localidades. Os resultados são mostrados na seção 4.5.1.

As funções de covariância espaço-temporal separável (4.2) e não-separável (4.3) da família de Gneiting (2002) (ver p. 69) também foram ajustadas aos dados para predizer a taxa de criminalidade nas 5 localidades de validação nos três últimos anos (1995 a 1997) e os resultados deste ajuste são apresentados na seção 4.5.2.

4.5.1 Combinação dos Modelos de Geoestatística e de Séries Temporais

O modelo de Høst et al. (1995) (ver seção 3.5.1, p. 53) foi ajustado aos dados na primeira etapa de predição considerando-se o logaritmo da taxa de criminalidade. A componente *F* foi calculada e pelo método de máxima verossimilhança o variograma teórico ajustado a esta componente foi o cúbico e os parâmetros estimados para a média, efeito pepita, variância e escala, são respectivamente: 4,31; 0,015; 0,15 e 2,95, ou seja:

$$\gamma(h) = 0{,}015 + 0{,}15\begin{cases} 7\left(\frac{h}{2{,}95}\right)^2 - 8{,}75\left(\frac{h}{2{,}95}\right)^3 + 3{,}5\left(\frac{h}{2{,}95}\right)^5 - 0{,}75\left(\frac{h}{2{,}95}\right)^7, & \text{se } h < 2{,}95 \\ 0, & \text{C.C.} \end{cases} \quad (4.11)$$

O variograma teórico ajustado à componente *F* é apresentado na Figura 4.20.

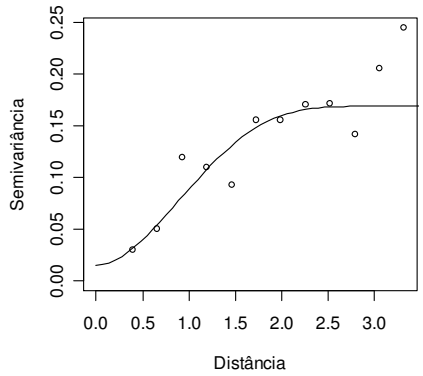


Figura 4.20. Variograma teórico ajustado à componente *F*.

A quantidade de vizinhos utilizada na predição foi determinada pelo menor valor do erro quadrático médio (EQM) e pela distância *di*. Este erro é baseado nas observações reais das 5 localidades de validação no período de 1986 a 1994. Pelo critério do EQM considerou-se na predição somente o vizinho mais próximo. O parâmetro de alcance estimado pelo modelo cúbico ajustado a componente *F* é igual a 2,95 e este valor foi utilizado como a distância *di* para quantificar a vizinhança. Pela distância *di* o número de vizinhos usados nas predições nas regiões Alto do Jequitinhonha, Central, Noroeste, Vale do Rio Grande e Vale do Sapucaí são respectivamente: 11, 14, 3, 5 e 8.

A Figura 4.21 mostra o EQM de acordo com o número de vizinhos. A quantidade de vizinhos igual a 1 corresponde ao menor valor do EQM. A Figura 4.22 mostra o EQM de acordo com a distância. A distância *di* igual a 2,95 graus conforme sugerido pelo variograma dado na equação (4.11) é assinalada por "*" no gráfico.

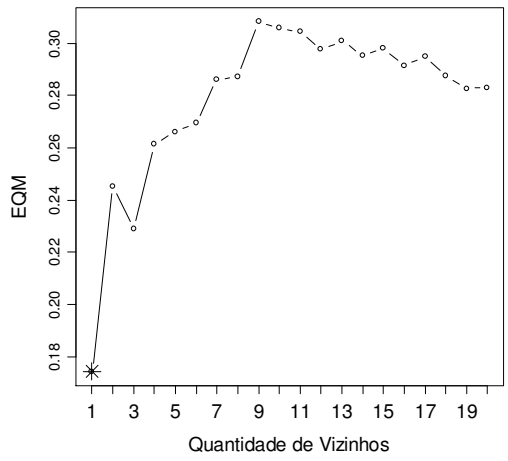


Figura 4.21. EQM de acordo com a quantidade de vizinhos.

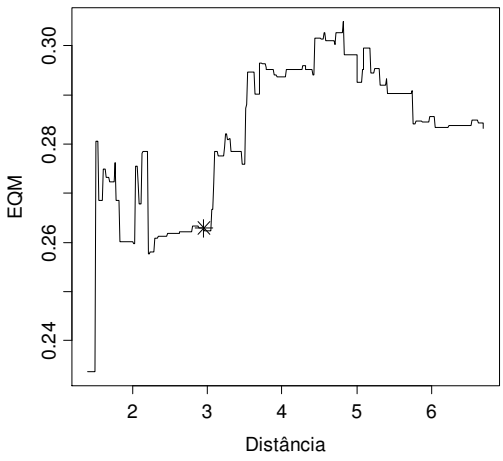


Figura 4.22. EQM de acordo com a distância.

Denotaremos estes modelos de Høst et al. (1995) ajustados aos dados pelos critérios do EQM e da distância *di* de respectivamente: Modelo 7 e Modelo 8.

Pela validação cruzada o ajuste dos Modelos 7 e 8 aos dados foi considerado adequado. Os valores da média e do desvio–padrão obtidos pela validação para os dois modelos são iguais a respectivamente: (0,0171; 0,4628) e (0,0386; 0,3785). O primeiro valor dentro dos parêntesis corresponde à média e o segundo é referente ao desvio–padrão. Os resíduos para ambos os modelos tem uma distribuição aproximadamente normal.

A Tabela 4.9 apresenta os resultados dos ajustes pelos Modelos 7 e 8.

Tabela 4.9 – Comparação dos modelos 7 e 8.

Modelo	EQM	RES
Høst et al. (1995): EQM – Modelo 7	0,1744	0,1480
Høst et al. (1995): *di* – Modelo 8	0,2629	0,2192

Na etapa 2 de predição o banco de dados original é atualizado com as predições obtidas para as localidades de validação pelo ajuste dos Modelos 7 e 8 determinando dois novos bancos de dados: banco 1 e banco 2. A estas bases de dados ajustamos o modelo baseado na proposta de Niu et al. (2003) (ver seção 3.5, p. 54), porém para o banco 1 utilizamos o critério do EQM para o cálculo do número de vizinhos da equação (3.44) (ver p. 54) e para o banco 2 usamos o critério da distância *di.* Os critérios adotados para cada banco na etapa 2 correspondem aos mesmos critérios empregados na etapa 1 de predição.

O modelo ajustado aos dados do banco 1 baseado na proposta de Niu et al. (2003) que utiliza o critério do EQM é dado em (4.12).

$$\begin{aligned} Z(s,t) = \; & 0{,}1033 + 0{,}7567 \times Z(s,t-1) + 0{,}0533 \times Z(v_1,t-1) - 0{,}0730 \times Z(v_2,t-1) \\ & + 0{,}0370 \times Z(v_3,t-1) + \ldots - 0{,}0003 \times Z(v_{23},t-1) + 0{,}2152 \times Z(v_{24},t-1), \\ & t = 1995,1996,1997 \end{aligned} \qquad (4.12)$$

A equação (4.12) diz que a predição do log da taxa em cada uma das cinco localidades de validação nos anos de 1995, 1996 e 1997 utiliza as informações de toda a vizinhança e do próprio local de predição no ano anterior ao da previsão adicionada de uma constante igual a 0,1033. Observamos que a contribuição de alguns vizinhos na predição é muito pequena.

Os resultados da predição pelo ajuste do modelo (4.12) de acordo com o ano são mostrados na Tabela 4.10. Os erros nesta tabela são calculados usando as predições do log da taxa nas cinco localidades de validação em cada ano.

Tabela 4.10 – Resultados da predição pelo ajuste do modelo (4.12).

Tempo	EQM	RES	MSQE
1995	0,2157	0,0760	4,395
1996	0,2124	0,1366	----
1997	0,2552	0,2256	----

A segunda maneira de estimar as observações no período considerado é baseada no critério da distância *di*. Para isto, calculou-se o variograma amostral usando os dados no ano de 1994. O variograma teórico ajustado aos dados foi o gaussiano e os parâmetros estimados para a média, efeito pepita, variância e escala, são respectivamente: 4,38; 0; 0,2308 e 1,2016, i.e.:

$$\gamma(h) = \begin{cases} 0{,}2308\left\{1 - \exp\left[-\left(\dfrac{h}{1{,}2016}\right)^2\right]\right\} & , h > 0 \\ 0 & , h \leq 0 \end{cases} \tag{4.13}$$

A distância *di* foi determinada pelo alcance prático, pois o modelo gaussiano atinge o patamar assintoticamente. O alcance prático é igual ao parâmetro de alcance estimado pelo modelo multiplicado por uma constante igual a 1,73. Dessa forma temos que $di = 1{,}73 \times 1{,}2016 = 2{,}08$ graus. A quantidade de vizinhos pelo critério da distância *di* é igual a 5.

O modelo ajustado aos dados do banco 2 baseado no critério da distância *di* é dado em (4.14).

$$\begin{aligned} Z(s,t) = \; & 0{,}4575 + 0{,}8990 \times Z(s,t-1) + 0{,}0513 \times Z(v_1,t-1) - 0{,}0615 \times Z(v_2,t-1) \\ & + 0{,}0040 \times Z(v_3,t-1) + 0{,}0142 \times Z(v_4,t-1) - 0{,}0085 \times Z(v_5,t-1), \\ & t = 1995,1996,1997 \end{aligned} \tag{4.14}$$

Os resultados da predição pelo ajuste do modelo (4.14) de acordo com o ano são mostrados na Tabela 4.11.

Tabela 4.11 – Resultados da predição pelo ajuste do modelo (4.14).

Tempo	EQM	RES	MSQE
1995	0,2696	0,1482	5,1481
1996	0,2990	0,2327	----
1997	0,4014	0,3391	----

Os modelos (4.12) e (4.14) ajustados aos dados são denotados por respectivamente: Modelos 9 e 10.

Observamos pelas Tabelas (4.10) e (4.11) que o Modelo 9 apresenta erros menores de predição se comparado com o Modelo 10, sendo que a diferença é maior no ano de 1997. Porém, o Modelo 9 utiliza toda a vizinhança na predição das observações enquanto o modelo 10 usa a informação somente dos 5 vizinhos mais próximos (ver equações (4.12) e (4.14)).

O ajuste pelos Modelos 9 e 10 serão comparados na seção 4.5.3 com o ajuste dos dados pelas funções de covariância da família de Gneiting (2002). A seção seguinte apresenta o ajuste por estas funções.

4.5.2 Ajuste por Funções de Covariância da Família de Gneiting (2002)

As funções de covariância espaço–temporal separável (4.2) e não-separável (4.3) da família de Gneiting (ver p. 69) estimadas pelo método de máxima verossimilhança utilizando os dados do período de 1986 a 1994 e as 20 localizações são dadas respectivamente por (4.15) e (4.16):

$$C(h,u)=\frac{2{,}08}{\left(0{,}1254|u|^2+1\right)^{0{,}35}}\exp\left(-0{,}0146\|h\|^{0{,}60}\right)+\frac{0{,}0143}{\left(0{,}1254|u|^2+1\right)^{0{,}35}} \tag{4.15}$$

$$C(h,u)=\left(\frac{2{,}08}{\left(0{,}1254|u|^2+1\right)^{0{,}0168}}\exp\left(-0{,}0146\left[\frac{\|h\|}{\left(0{,}1254|u|^2+1\right)^{0{,}0084}}\right]^{0{,}60}\right)\right)\times\frac{1}{\left(0{,}1254|u|^2+1\right)^{0{,}35}}+\frac{0{,}0143}{\left(0{,}1254|u|^2+1\right)^{0{,}35}} \tag{4.16}$$

O parâmetro β estimado em (4.16) é igual a 0,0168, o que indica uma interação fraca entre os processos espacial e temporal.

A verossimilhança condicional e a perfilhada são mostradas na Figura 4.23. A verossimilhança condicional está condizente com o resultado anterior indicando que o modelo separável é mais adequado. A verossimilhança perfilhada apresenta o máximo em $\beta=0{,}2$ e este valor é próximo de quando $\beta=0$. O estudo das verossimilhanças confirma neste caso também a instabilidade e a dificuldade de estimação do parâmetro β.

A predição nas 5 localidades de validação nos anos de 1995, 1996 e 1997 será calculada usando o modelo separável dado em (4.15) e denotaremos o ajuste por Modelo 11.

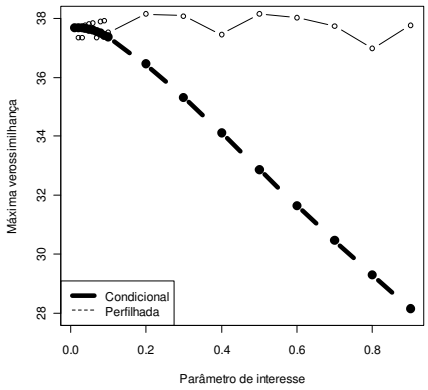


Figura 4.23. Máxima verossimilhança condicional e perfilhada.

A comparação dos ajustes dos dados pelos Modelos 9, 10 e 11 é mostrada na seção seguinte.

4.5.3 Comparação dos Modelos Ajustados: Modelos 9, 10 e 11

A Tabela 4.12 compara as predições utilizando os Modelos 9, 10 e 11. Os Modelos 9 e 10 foram obtidos pela combinação das metodologias de geoestatística e de séries temporais, sendo que o primeiro utiliza o EQM para o cálculo do número de vizinhos e o segundo é baseado no critério da distância *di*. O Modelo 11 é concernente a função de covariância separável da família de Gneiting (2002) dada em (4.15). Os erros são calculados pelas predições do log da taxa feitas nas 5 localidades de validação nos três instantes de tempo (1996, 1997 e 1998).

Tabela 4.12 - Comparação dos modelos 9, 10 e 11.

Modelo	EQM	RES
Combinado: EQM - Modelo 9	0,2278	0,1460
Combinado: *di* - Modelo 10	0,3233	0,2400
Gneiting (2002): Separável - Modelo 11	0,5827	0,3941

Os Modelos 9 e 10 apresentam os melhores ajustes se comparados com o Modelo 11. A Figura 4.24 apresenta o EQM para o log da taxa de acordo com a região

administrativa. Observamos que a região Central é a que apresenta os maiores erros de predição e os menores erros correspondem às regiões do Vale do Rio Grande e do Vale do Sapucaí considerando os três modelos.

A Figura 4.25 mostra o EQM de acordo com o ano. Os erros associados as predições no ano de 1997 são maiores que para os outros anos para os três modelos ajustados. Isto ocorre pelo fato de que as predições no ano de 1997 são baseadas nas predições dos anos de 1995 e 1996.

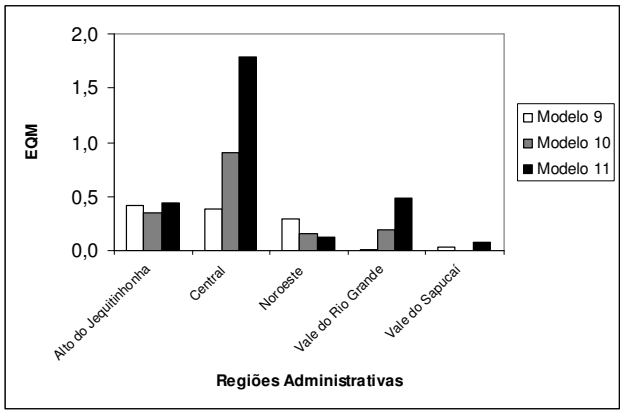


Figura 4.24. EQM de acordo com a região administrativa.

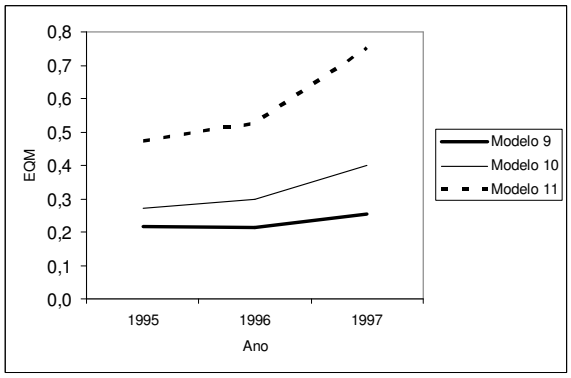
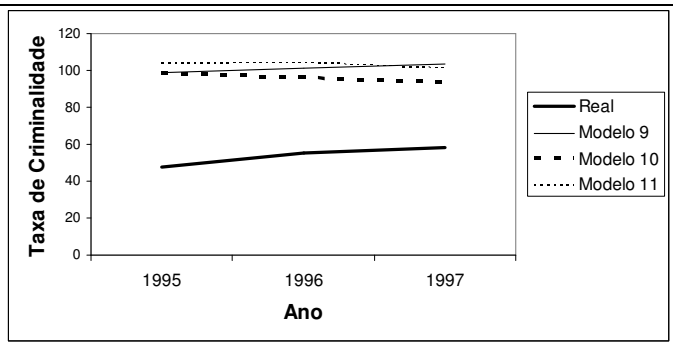


Figura 4.25. EQM de acordo com o ano.

Os valores preditos pelos Modelos 9, 10 e 11 na escala original e os valores observados para as cinco regiões de validação são apresentadas nos Quadros de 4.11 a 4.15. As melhores predições correspondem às regiões do Vale do Rio Grande e do Vale do Sapucaí.

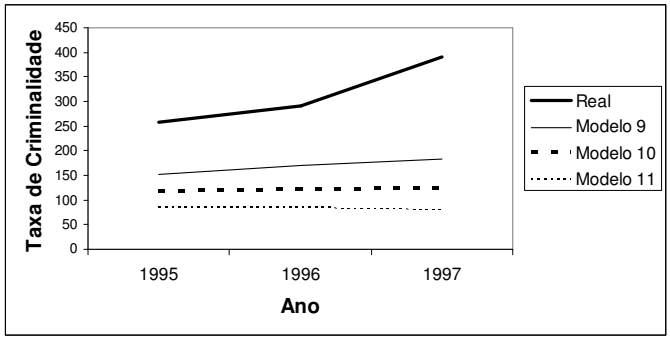
Quadro 4.11. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa Alto do Jequitinhonha.

Alto do Jequitinhonha				
Ano	Observado	Modelo 9	Modelo 10	Modelo 11
1995	47,56	98,90	98,12	103,51
1996	55,22	101,23	95,60	104,28
1997	58,00	103,81	93,55	100,95
Erro médio		-47,72	-42,16	-49,32
$\sqrt{\text{EQM}}$		47,79	42,62	49,61



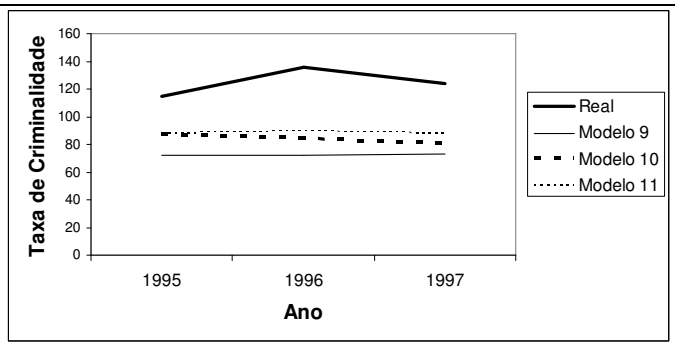
Quadro 4.12. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa Central.

Central				
Ano	Observado	Modelo 9	Modelo 10	Modelo 11
1995	258,94	151,90	117,39	83,96
1996	290,10	169,27	120,50	83,07
1997	389,46	183,60	122,95	79,58
Erro médio		144,58	192,55	230,63
$\sqrt{EQM}$		151,04	199,86	237,70



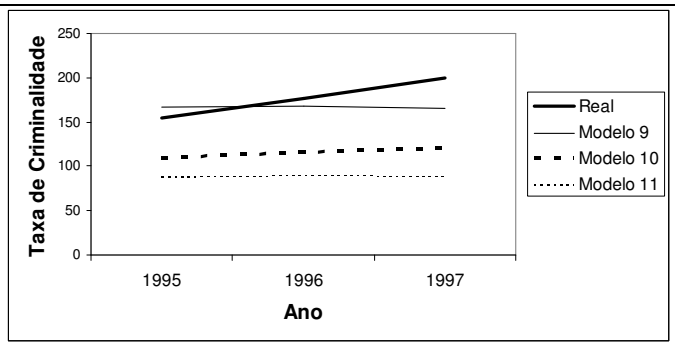
Quadro 4.13. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa Noroeste.

Noroeste				
Ano	Observado	Modelo 9	Modelo 10	Modelo 11
1995	114,65	72,19	86,84	87,63
1996	135,89	72,36	83,67	89,57
1997	123,57	72,97	81,06	87,85
Erro médio		52,20	40,85	36,35
$\sqrt{EQM}$		52,91	42,06	37,20



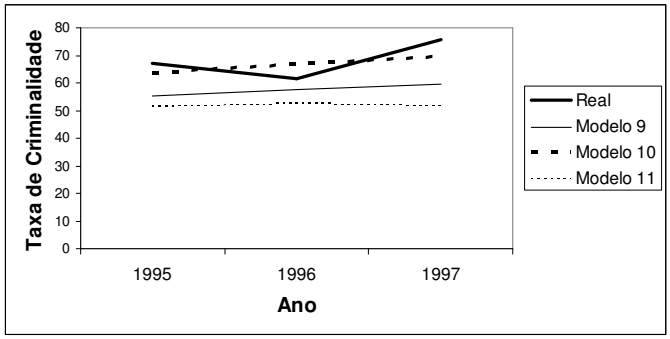
Quadro 4.14. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa do Vale do Rio Grande.

Vale do Rio Grande				
Ano	Observado	Modelo 9	Modelo 10	Modelo 11
1995	154,02	166,60	109,30	86,48
1996	176,07	167,34	115,09	89,28
1997	200,17	165,29	119,71	88,24
Erro médio		10,34	62,05	88,75
$\sqrt{EQM}$		21,99	63,75	90,60



Quadro 4.15. Valores observados e preditos da taxa de criminalidade por ano para a região administrativa do Vale do Sapucaí.

Vale do Sapucaí				
Ano	Observado	Modelo 9	Modelo 10	Modelo 11
1995	67,18	55,30	63,69	51,23
1996	61,67	57,55	66,55	52,40
1997	75,60	59,47	69,44	51,75
Erro médio		10,71	1,59	16,36
$\sqrt{\text{EQM}}$		11,81	4,96	17,41



Os erros globais das predições na escala original da variável taxa de criminalidade são resumidos na Tabela 4.13 de acordo com o modelo utilizado na predição.

Tabela 4.13 – Análise descritiva dos erros.

Método	Média	Média absoluta	Desvio padrão	Mínimo	Máximo
Combinação: EQM – Modelo 9	34,02	54,79	69,79	-51,34	205,86
Combinação: *di* – Modelo 10	50,98	68,49	86,14	-50,56	266,51
Gneiting (2002): Separável – Modelo 11	64,55	84,28	101,44	-55,95	309,88

O Modelo 9 é o que apresenta os menores erros médios de predição se comparado com os Modelos 10 e 11. A variabilidade de predição é alta para todos os modelos considerados.

Pelas análises apresentadas neste capítulo podemos concluir que os modelos modificados de Høst et al. (1995) e de Niu et al. (2003) usando o critério da distância *di* são alternativas razoáveis para estimar a taxa de criminalidade nas regiões administrativas do estado de MG quando o interesse é respectivamente a interpolação espacial em tempos observados na amostra ou a previsão temporal em localizações amostradas. Estes modelos são preferíveis ao ajuste pelas funções de covariância de Gneiting (2002), visto que estas acarretam maiores erros nas predições.

Os resultados da predição da taxa de criminalidade em localizações e instantes de tempo não observados na amostra utilizando a combinação dos modelos de Høst et al. (1995) e de Niu et al. (2003) foram melhores que aqueles obtidos aplicando a função de covariância separável da classe de Gneiting (2002).

Para todos os três casos de análise observamos que os erros de predição são menores nos locais onde os vizinhos apresentam um comportamento semelhante da característica de interesse e as séries de dados são mais estáveis.

Capítulo 5

Estudo de Caso: Armazenagem de Água em um Solo Cultivado com Citros

Neste capítulo ajustamos os modelos apresentados no capítulo 3 aos dados de armazenagem de água em um solo cultivado com citros[11].

5.1 Introdução

Os dados de armazenagem de água foram ajustados pelo modelo de Høst et al. (1995) com as modificações apresentadas previamente (ver seção 3.5.1, p. 53), pelo modelo baseado nas idéias de Niu et al. (2003) (ver seção 3.5.2, p. 54), pela combinação destes modelos (ver seção 3.6, p. 56) e por funções de covariância espaço-temporal separável e não-separável da família de Gneiting (2002) (ver seção 4.3.2, p. 69). O ajuste dos modelos aos dados é muito similar a aquele apresentado no capítulo 4 e em razão da semelhança destes procedimentos, algumas partes da exposição serão omitidas. A notação adotada para a identificação dos modelos ajustados é a mesma que a apresentada no capítulo 4.

A seção seguinte descreve os dados utilizados neste capítulo.

5.2 Descrição dos Dados

Os dados[12] se referem à armazenagem de água em um solo classificado como Latossolo Vermelho Amarelo Argissólico cultivado com citros. As amostras foram coletadas em 40 pontos que são distribuídos em duas transeções (ou linhas) com 20 pontos de observação cada. A distância entre as plantas é igual a 4,0 m e as transeções são separadas por uma distância igual a 7,0 m. A distribuição dos locais de coleta dos dados pode ser

[11] Os citros compreendem um grande grupo de plantas do gênero Citrus e outros gêneros afins (Fortunella e Poncirus) ou híbridos da família Rutaceae, representado, na maioria, por laranjas (Citrus sinensis), tangerinas (Citrus reticulata e Citrus deliciosa), limões (Citrus limon), limas ácidas como o Tahiti (Citrus latifolia) e o Galego (Citrus aurantiifolia), e doces como a lima da Pérsia (Citrus limettioides), pomelo (Citrus paradisi), cidra (Citrus medica), laranja-azeda (Citrus aurantium) e toranjas (Citrus grandis).
(http://www.iac.sp.gov.br/Tecnologias/Citros/Citros.htm, acesso em 11/2007).

[12] Estes dados fazem parte do trabalho de pesquisa de Moreti (2006) que realiza um estudo da armazenagem de água neste tipo de solo.

visualizada na Figura 5.1. Os oito pontos (ou 20% dos dados) marcados por "*" foram os locais separados para validação dos modelos e a seleção destas localizações foi aleatória. Estes pontos são identificados pela numeração de 1 a 8 na análise de ajuste de modelos, sendo que 1 corresponde ao primeiro ponto "*" da esquerda para a direita na linha inferior, e o 8 corresponde ao último ponto "*" da esquerda para a direita na linha superior.

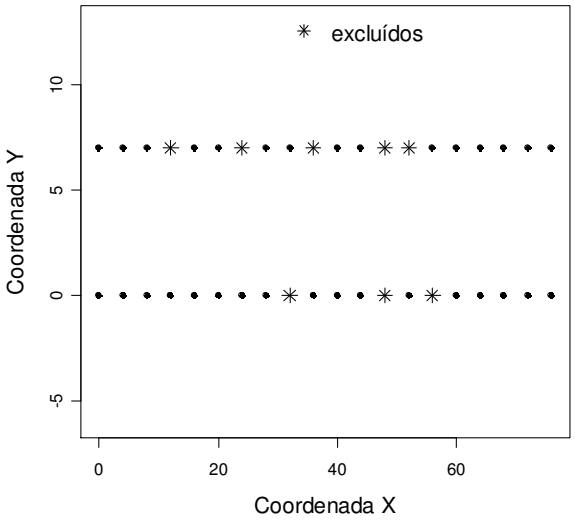


Figura 5.1. Distribuição dos locais de coleta dos dados.

A coleta dos dados foi feita a partir da instalação de um tubo de acesso a uma sonda de nêutrons a uma profundidade de até 1,20 m em cada um dos pontos amostrais (MORETI, 2006). Os dados foram observados em 98 semanas não eqüidistantes no tempo ao longo de três anos (2001-2004). Nesta dissertação utilizamos apenas 23 instantes de tempo devido a problemas computacionais no ajuste do modelo de covariância da família de Gneiting (2002), sendo assim tomamos um instante de tempo a cada quatro para compor a amostra. O leitor interessado em outras informações a respeito destes dados deve consultar o trabalho de Moreti (2006).

A Tabela 5.1 mostra a média e o desvio-padrão da armazenagem de água nos 23 instantes de tempo para os 40 pontos e cada um dos pontos é localizado pelas coordenadas x e y. Os pontos separados para a validação de número 1 a 8 correspondem na Tabela 5.1 as seguintes observações respectivamente: 9, 13, 15, 24, 27, 30, 33 e 34.

Tabela 5.1 – Análise Descritiva dos Dados de Armazenagem de Água.

Observação	x	y	Média	Desvio padrão	Observação	x	y	Média	Desvio padrão
1	0	0	17,46	2,56	21	0	7	17,85	2,3
2	4	0	16,76	2,26	22	4	7	17,19	2,29
3	8	0	18,01	2,3	23	8	7	16,89	2,35
4	12	0	18,21	2,07	* 24	12	7	17,49	2,09
5	16	0	19,05	2,37	25	16	7	17,62	2,26
6	20	0	17,09	2,04	26	20	7	16,9	2,04
7	24	0	17,11	1,9	* 27	24	7	16,24	2,09
8	28	0	16,4	2,18	28	28	7	16,54	1,98
* 9	32	0	17,09	2	29	32	7	17,04	2,04
10	36	0	17,66	1,76	* 30	36	7	16,71	2,13
11	40	0	16,25	2,05	31	40	7	17,16	2,19
12	44	0	16,67	2,01	32	44	7	17,72	2,09
* 13	48	0	17,03	2,06	* 33	48	7	16,92	1,85
14	52	0	17,05	2,12	* 34	52	7	16,98	1,79
* 15	56	0	15,38	1,84	35	56	7	16,83	1,94
16	60	0	17,19	1,85	36	60	7	16,94	2,01
17	64	0	15,64	1,71	37	64	7	16,37	2,32
18	68	0	16,21	1,81	38	68	7	17,14	1,94
19	72	0	17,48	2,1	39	72	7	16,85	1,99
20	76	0	17,15	1,85	40	76	7	16,92	1,78

* localizações separadas para validação.

Observamos pela Tabela 5.1 que a média de armazenagem de água no solo com citros nos 23 períodos considerados é semelhante entre os pontos amostrados e a variabilidade em cada um dos pontos é pequena. O menor valor médio da variável é referente a observação de número 15 que corresponde ao ponto separado da análise identificado como 3 no gráfico da Figura 5.1. A Figura 5.2 mostra o histograma da média da variável nos 23 instantes de tempo para cada uma das 40 localizações. A distribuição da média é simétrica e aproximadamente normal.

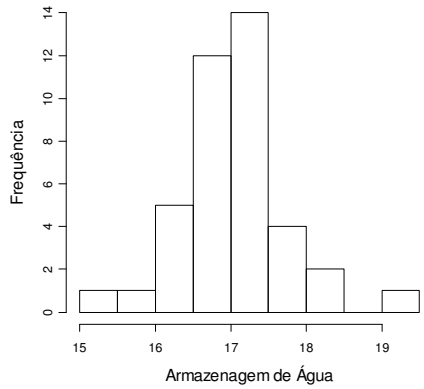


Figura 5.2. Histograma da média da variável nos 23 tempos em cada localização.

A Tabela 5.2 mostra os modelos ajustados a cada um dos casos de análise e as informações disponíveis na base de dados para a predição.

Tabela 5.2 – Modelos Ajustados a cada um dos Casos de Análise.

Caso	Modelo	Base de Dados		Predição	
		Localizações	Tempo	Localizações	Tempo
1	Høst et al. (1995): EQM - Modelo 1	32	23 (1-23)	8	23 (1-23)
	Høst et al. (1995): *di* - Modelo 2	32	23 (1-23)	8	23 (1-23)
	Gneiting (2002) – Modelo 3	32	23 (1-23)	8	23 (1-23)
2	Niu et al. (2003): EQM - Modelo 4	40	19 (1-19)	8	4 (20-23)
	Niu et al. (2003): *di* - Modelo 5	40	19 (1-19)	8	4 (20-23)
	Gneiting (2002) – Modelo 6	40	19 (1-19)	8	4 (20-23)
3	Høst et al. (1995): EQM - Modelo 7	32	19 (1-19)	8	19 (1-19)
	Høst et al. (1995): *di* - Modelo 8	32	19 (1-19)	8	19 (1-19)
	Niu et al. (2003): EQM - Modelo 9	40	19 (1-19)	8	4 (20-23)
	Niu et al. (2003): *di* - Modelo 10	40	19 (1-19)	8	4 (20-23)
	Gneiting (2002) – Modelo 11	40	19 (1-19)	8	4 (20-23)

A próxima seção apresenta o Caso 1 de análise.

5.3 Análise: Caso 1

No Caso 1 de análise o objetivo é predizer a variável armazenagem de água nos 8 pontos que foram separados do banco de dados original para testar o modelo nos 23 instantes de tempo. Esta predição é feita utilizando uma base de dados formada por 32 pontos amostrais e para cada ponto tem-se a informação da armazenagem de água nos 23 tempos.

Uma análise exploratória dos dados foi realizada inicialmente com o objetivo de se conhecer melhor os dados e auxiliar na implementação dos modelos. A distribuição dos dados é aproximadamente normal e não foi feita nenhuma transformação nos dados. A Figura 5.3 mostra a armazenagem de água nos 32 pontos ao longo do tempo e podemos observar que o comportamento temporal da variável é semelhante entre os pontos amostrados.

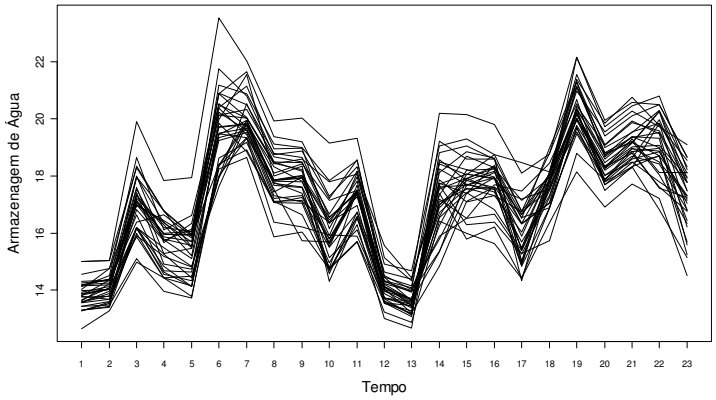


Figura 5.3. Série temporal para cada uma das 32 localidades.

Para cada um dos tempos separadamente ajustamos um modelo geoestatístico pelo método de máxima verossimilhança, ainda como uma análise descritiva dos dados. O modelo de variograma ajustado aos dados para todos os anos é o exponencial[13] como mostra a Figura 5.4.

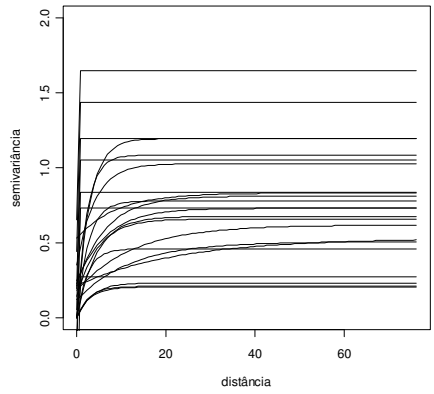


Figura 5.4. Variogramas ajustados para cada tempo separadamente.

Os parâmetros estimados do variograma para cada tempo podem ser visualizados na Figura 5.5. A linha horizontal no gráfico corresponde a média das estimativas nos 23 tempos observados. Aparentemente não existe tendência nas componentes estimadas de média (a), de efeito pepita (b), de variância (c) e de escala (d) ao longo do tempo.

[13] Em alguns tempos o modelo ajustado foi diferente do exponencial, mas como os valores do AIC, BIC e o logaritmo da verossimilhança eram muito similares entre o modelo inicialmente ajustado e o modelo exponencial, adotamos o último para que fosse possível avaliar a evolução das estimativas dos parâmetros do modelo ao longo do tempo.

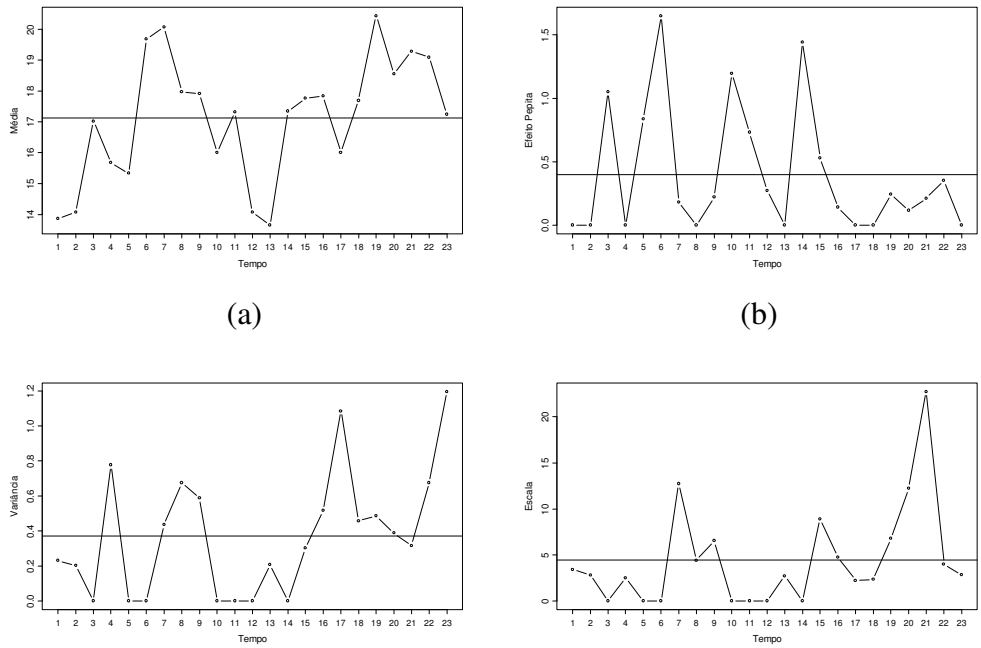


Figura 5.5. Parâmetros estimados para a média (a), efeito pepita (b), variância (c) e escala (d).

O ajuste dos dados pelo modelo proposto por Høst et al. (1995) é apresentado na seção seguinte. Na seção 5.3.2 mostramos o ajuste por funções de covariância da família de Gneiting (2002). Os resultados dos ajustes são comparados na seção 5.3.3.

5.3.1 Ajuste pelo Modelo Proposto por Høst et al. (1995)

Nesta seção mostramos o ajuste dos dados pelo modelo proposto por Høst et al. (1995) com os parâmetros estimados pelo método de Kyriakidis e Journel (1999) e com as modificações discutidas no Capítulo 3.

O modelo cúbico foi ajustado à componente *F* e os parâmetros estimados para a média, o efeito pepita, a variância e a escala são respectivamente: 17,12; 0,23; 0,18 e 15,04, i.e.:

$$\gamma(h) = 0{,}23 + 0{,}18\begin{cases} 7\left(\frac{h}{15{,}04}\right)^2 - 8{,}75\left(\frac{h}{15{,}04}\right)^3 + 3{,}5\left(\frac{h}{15{,}04}\right)^5 - 0{,}75\left(\frac{h}{15{,}04}\right)^7, & \text{se } h < 15{,}04 \\ 0, & \text{caso contrário} \end{cases} \quad (5.1)$$

O variograma ajustado à componente *F* é apresentado na Figura 5.6.

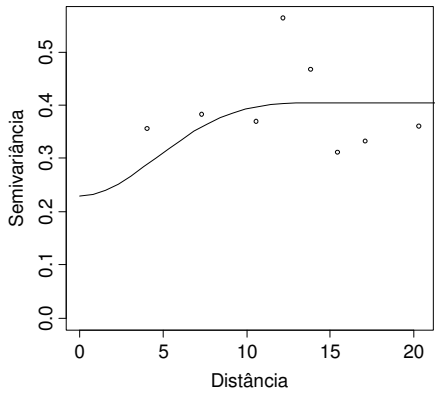


Figura 5.6. Variograma ajustado à componente *F*.

A predição da armazenagem de água nas 8 localidades separadas da base de dados original para testar a adequação do modelo nos 23 instantes de tempo foi calculada variando-se a quantidade de vizinhos utilizando os critérios do EQM e a distância *di*. Pelo EQM obtemos uma vizinhança formada pelos 7 pontos mais próximos do local de predição e a quantidade de vizinhos baseada na distância *di* para os pontos de 1 a 8 foram iguais respectivamente a: 11, 9, 10, 12, 10, 9, 9 e 10.

A distância *di* considerada no ajuste é igual a 15,04 m e corresponde ao parâmetro de alcance estimado pelo modelo de variograma cúbico ajustado à componente *F* de acordo com a equação (5.1). Os modelos baseados no EQM e na distância *di* são denotados por: Modelo 1 e Modelo 2.

A Figura 5.7 apresenta o EQM calculado pelo Modelo 1 de acordo com o número de vizinhos e a Figura 5.8 mostra a distância e o valor do EQM correspondente obtido pelo ajuste do Modelo 2. Observamos que a escolha de *di* igual a 15,04 m, assinalada no gráfico por "*" é razoável, pois resulta em um valor pequeno de erro se comparado com as outras distâncias. Estes erros foram computados utilizando as predições nas 8 localidades de validação nos 23 instantes de tempo.

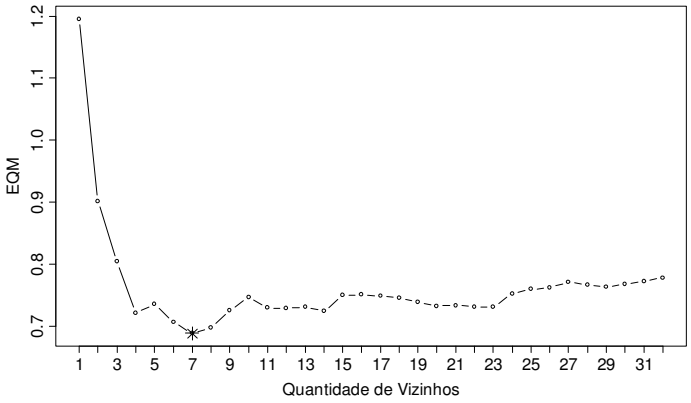


Figura 5.7. EQM de acordo com a quantidade de vizinhos.

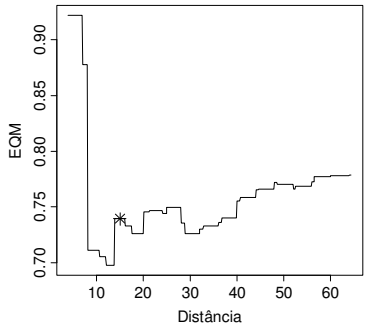


Figura 5.8. EQM de acordo com a distância.

Pela validação cruzada concluímos que os Modelos 1 e 2 se ajustaram adequadamente aos dados, pois a média dos resíduos para ambos os modelos foi próxima de zero (0,012; 0,016) e a distribuição dos resíduos é aproximadamente normal. A validação cruzada foi feita retirando-se um a um dos 32 pontos e estimando-os nos 23 instantes de tempo pelos Modelos 1 e 2, respectivamente.

A próxima seção apresenta o ajuste dos dados pelas funções de covariância da família de Gneiting (2002).

5.3.2 Ajuste por Funções de Covariância da Família de Gneiting (2002)

A função de covariância espaço–temporal separável dada pela equação (4.2) da família de Gneiting (2002) ajustada pelo método de máxima verossimilhança utilizando os dados nas 32 localidades e os 23 instantes de tempo é dada por (5.2):

$$C(h,u) = \frac{26{,}57}{\left(0{,}2239|u|^{0{,}9867}+1\right)^{0{,}2223}} \exp\left(-0{,}0013\|h\|^{0{,}8179}\right) + \frac{0{,}2517}{\left(0{,}2239|u|^{0{,}9867}+1\right)^{0{,}2223}} \qquad (5.2)$$

O valor do parâmetro β estimado pelo modelo não-separável da forma (4.3) ajustado aos dados de armazenagem de água é igual à zero. Dessa forma, o modelo não-separável se reduz ao modelo separável dado pela equação (5.2). O ajuste pela função dada em (5.2) é denotado como Modelo 3.

A Figura 5.9 apresenta o estudo da verossimilhança condicional e perfilhada onde confirmamos a adequação do modelo separável, visto que o máximo destas funções ocorre quando $\beta = 0$.

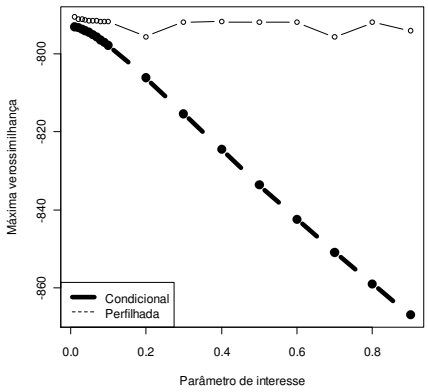


Figura 5.9. Máxima verossimilhança condicional e perfilhada.

A seção seguinte compara os ajustes pelos Modelos 1, 2 e 3.

5.3.3 Comparação dos Modelos Ajustados: Modelos 1, 2 e 3

A Tabela 5.3 mostra o erro quadrático médio (EQM) e a média dos resíduos (RES) pelo ajuste dos Modelos 1, 2 e 3 aos dados. Os Modelos 1 e 2 se referem à proposta de Høst et al. (1995) que utiliza respectivamente, o critério do erro quadrático médio (EQM) e o critério da distância *di* para a escolha do número de vizinhos para a predição, e o Modelo 3 é concernente à função de covariância separável pertencente à família de Gneiting dada na equação (5.2). Estes erros foram calculados usando as predições nas 8 localidades de validação nos 23 instantes de tempo.

Tabela 5.3 – Comparação dos modelos 1, 2 e 3.

Modelo	EQM	RES
Høst et al. (1995): EQM - Modelo 1	0,6882	-0,3439
Høst et al. (1995): *di* - Modelo 2	0,7394	-0,3567
Gneiting (2002): Separável - Modelo 3	0,7049	-0,3553

Os erros associados ao ajuste pelos Modelos 1, 2 e 3 são semelhantes. Aparentemente não existe diferença significativa entre os modelos ajustados.

A Figura 5.10 mostra o EQM de acordo com o ponto amostral de validação. O erro é calculado utilizando as predições nos 23 instantes de tempo para cada localização. Os erros são semelhantes em todos os pontos amostrais no que se refere ao modelo ajustado. O ponto

amostral 3 foi o que apresentou os maiores erros de predição se comparado com as outras localizações.

A Figura 5.11 mostra o EQM por tempo. O erro é calculado utilizando as predições nos 8 pontos amostrais de validação para cada tempo. Aparentemente não existe diferença entre os modelos ajustados no que se refere ao EQM, pois o comportamento temporal dos erros é semelhante entre os três modelos.

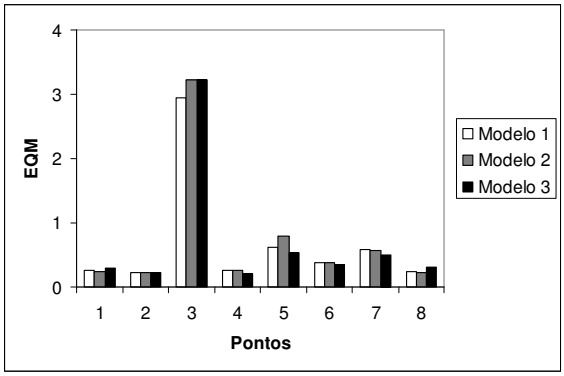


Figura 5.10. Erro quadrático médio por ponto de acordo com o modelo ajustado.

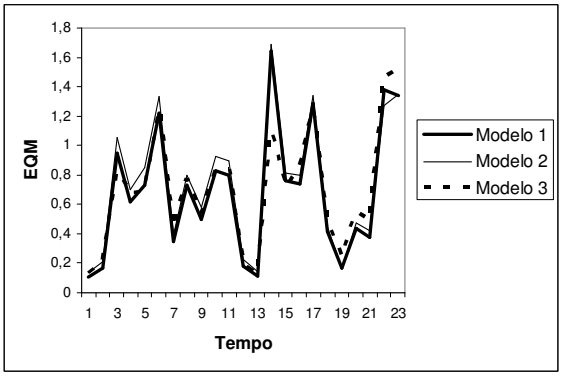


Figura 5.11. Erro quadrático médio por tempo de acordo com o modelo ajustado.

Os Quadros de 5.1 a 5.8 apresentam os valores observados e os valores preditos para os 8 pontos amostrais de validação nos 23 instantes de tempo considerando os três modelos descritos anteriormente para o ajuste. Observamos que os valores preditos pelo ajuste dos Modelos 1, 2 e 3 conseguem acompanhar a série de valores reais da armazenagem de água no período analisado em todos os pontos amostrais, exceto no ponto de validação 3. Neste ponto os valores preditos pelos três modelos superestimaram os valores reais (ver Quadro 5.3).

Se plotarmos os valores reais da armazenagem de água no ponto amostral 3 de validação ao longo do tempo conjuntamente com os valores dos 4 vizinhos mais próximos, observamos que a série de valores da variável no ponto 3 é inferior as séries construídas com as informações dos vizinhos, como mostra a Figura 5.12.

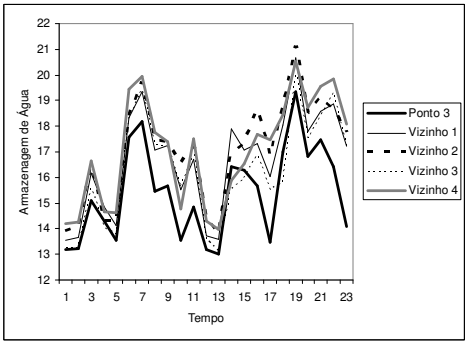
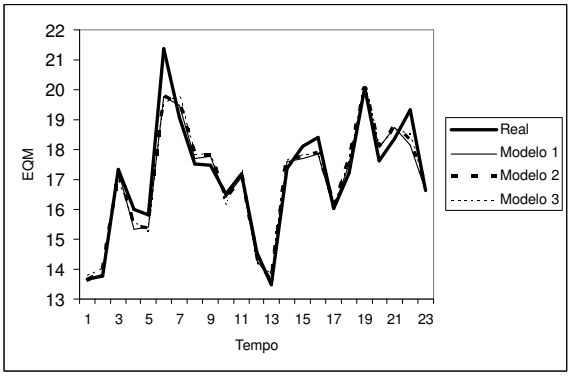


Figura 5.12. Armazenagem de água no ponto 3 e nos 4 vizinhos mais próximos ao longo do tempo.

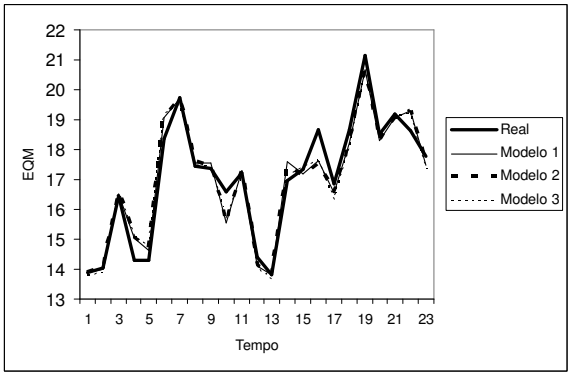
Quadro 5.1. Valores observados e preditos da armazenagem de água por tempo para o ponto 1.

Ponto 1				
Tempo	Real	Modelo 1	Modelo 2	Modelo 3
1	13,65	13,6	13,64	13,79
2	13,78	13,85	13,88	14,01
3	17,32	17,28	17,21	16,99
4	16,01	15,35	15,44	15,57
5	15,83	15,4	15,38	15,28
6	21,36	19,78	19,79	19,59
7	19,07	19,5	19,53	19,81
8	17,53	17,72	17,77	17,81
9	17,5	17,79	17,8	17,85
10	16,52	16,41	16,35	16,15
11	17,19	17,08	17,12	17,32
12	14,54	14,3	14,3	14,22
13	13,49	13,62	13,64	13,84
14	17,37	17,59	17,72	17,66
15	18,12	17,69	17,75	17,84
16	18,4	17,84	17,89	17,89
17	16,03	16,03	16,06	16,11
18	17,21	17,53	17,62	17,61
19	20,09	19,96	20,06	20,17
20	17,64	18,09	18,12	18,07
21	18,38	18,75	18,78	18,64
22	19,33	18,15	18,32	18,5
23	16,62	16,63	16,69	16,65
Erro médio		0,13	0,09	0,07
$\sqrt{EQM}$		0,51	0,49	0,53



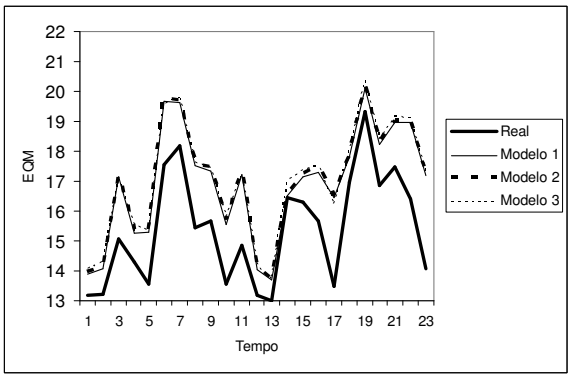
Quadro 5.2. Valores observados e preditos da armazenagem de água por tempo para o ponto 2.

Ponto 2				
Tempo	Real	Modelo 1	Modelo 2	Modelo 3
1	13,9	13,89	13,9	13,76
2	14,04	14,01	14,02	13,89
3	16,46	16,53	16,59	16,57
4	14,3	15,05	15,08	15,06
5	14,3	14,64	14,69	14,79
6	18,36	19,07	19,14	19,17
7	19,73	19,7	19,7	19,48
8	17,46	17,57	17,58	17,48
9	17,37	17,54	17,48	17,39
10	16,58	15,55	15,54	15,71
11	17,22	17,28	17,28	17,12
12	14,42	14,11	14,11	14,04
13	13,83	13,83	13,82	13,67
14	16,97	17,61	17,48	17,16
15	17,35	17,18	17,16	17,4
16	18,66	17,62	17,57	17,62
17	16,86	16,56	16,57	16,34
18	18,68	18,24	18,2	18,08
19	21,15	20,66	20,62	20,55
20	18,51	18,31	18,31	18,46
21	19,17	19,06	19,06	19,16
22	18,63	19,28	19,28	19,2
23	17,76	17,5	17,51	17,33
Erro médio		0,04	0,04	0,10
$\sqrt{EQM}$		0,47	0,48	0,47



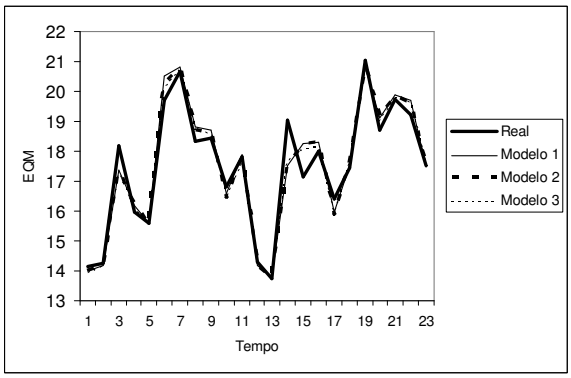
Quadro 5.3. Valores observados e preditos da armazenagem de água por tempo para o ponto 3.

Ponto 3				
Tempo	Real	Modelo 1	Modelo 2	Modelo 3
1	13,19	13,9	13,98	14,07
2	13,24	14,09	14,14	14,28
3	15,08	17,14	17,24	17,06
4	14,29	15,25	15,37	15,57
5	13,56	15,29	15,42	15,33
6	17,56	19,67	19,8	19,61
7	18,19	19,63	19,7	19,76
8	15,46	17,51	17,61	17,64
9	15,67	17,35	17,47	17,44
10	13,56	15,57	15,7	15,8
11	14,85	17,21	17,33	17,24
12	13,17	14,03	14,12	14,13
13	12,99	13,72	13,78	13,7
14	16,44	16,52	16,58	17,01
15	16,28	17,16	17,27	17,38
16	15,66	17,31	17,43	17,52
17	13,48	16,44	16,51	16,22
18	16,98	17,8	17,85	17,97
19	19,34	20,11	20,17	20,37
20	16,84	18,24	18,3	18,4
21	17,47	18,96	19,03	19,13
22	16,42	18,98	18,99	19,11
23	14,07	17,18	17,24	17,25
Erro médio		-1,53	-1,62	-1,66
$\sqrt{EQM}$		1,72	1,80	1,80



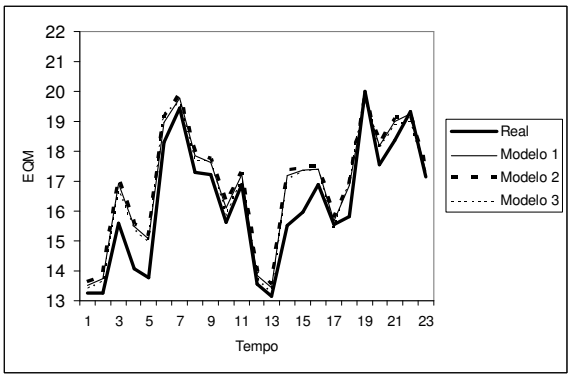
Quadro 5.4. Valores observados e preditos da armazenagem de água por tempo para o ponto 4.

Ponto 4				
Tempo	Real	Modelo 1	Modelo 2	Modelo 3
1	14,16	14,01	14	13,94
2	14,27	14,17	14,16	14,16
3	18,19	17,38	17,34	17,27
4	15,96	16,19	16,17	16,05
5	15,58	15,61	15,57	15,59
6	19,72	20,51	20,3	20,1
7	20,68	20,83	20,8	20,7
8	18,33	18,82	18,72	18,67
9	18,43	18,71	18,64	18,6
10	16,81	16,58	16,41	16,43
11	17,82	17,88	17,76	17,6
12	14,3	14,18	14,17	14,18
13	13,75	13,8	13,78	13,7
14	19,02	17,52	17,47	17,65
15	17,15	18,27	18,25	18,07
16	18,01	18,31	18,3	18,11
17	16,4	15,97	15,85	15,86
18	17,43	17,52	17,57	17,58
19	21,05	20,81	20,83	20,77
20	18,72	19,12	19,13	19,01
21	19,73	19,9	19,89	19,76
22	19,23	19,69	19,59	19,63
23	17,52	17,5	17,5	17,51
Erro médio		-0,04	0,00	0,06
$\sqrt{EQM}$		0,51	0,51	0,46



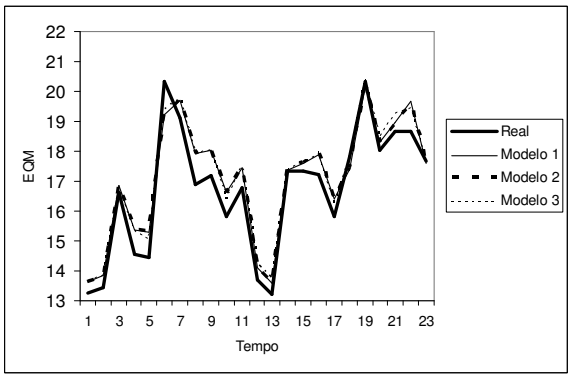
Quadro 5.5. Valores observados e preditos da armazenagem de água por tempo para o ponto 5.

Ponto 5				
Tempo	Real	Modelo 1	Modelo 2	Modelo 3
1	13,27	13,53	13,63	13,39
2	13,26	13,75	13,85	13,62
3	15,61	16,93	17,12	16,7
4	14,06	15,49	15,56	15,37
5	13,78	15,09	15,19	14,95
6	18,3	18,95	19,2	19,22
7	19,46	19,77	19,89	19,62
8	17,28	17,84	17,95	17,67
9	17,21	17,63	17,75	17,66
10	15,62	16,1	16,27	15,87
11	16,9	17,21	17,31	17,04
12	13,57	13,84	13,97	13,71
13	13,13	13,46	13,56	13,22
14	15,52	17,19	17,35	17,02
15	15,96	17,37	17,48	17,33
16	16,88	17,39	17,48	17,38
17	15,54	15,68	15,79	15,45
18	15,82	16,84	16,91	16,96
19	20,01	19,73	19,78	19,86
20	17,54	18,18	18,26	18,08
21	18,37	19,01	19,1	18,9
22	19,35	19,25	19,17	18,98
23	17,14	17,32	17,43	17,1
Erro médio		-0,61	-0,71	-0,50
$\sqrt{EQM}$		0,79	0,89	0,73



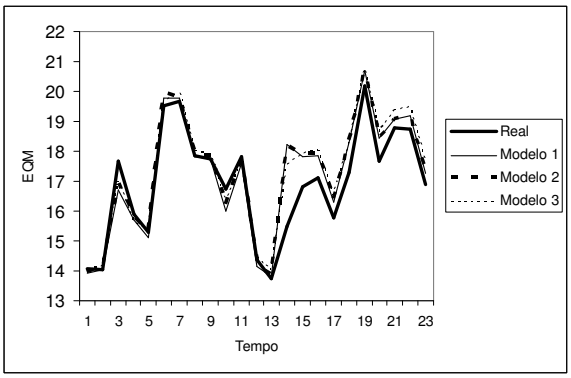
Quadro 5.6. Valores observados e preditos da armazenagem de água por tempo para o ponto 6.

Ponto 6				
Tempo	Real	Modelo 1	Modelo 2	Modelo 3
1	13,25	13,63	13,64	13,67
2	13,43	13,85	13,87	13,86
3	16,63	16,85	16,88	16,75
4	14,55	15,37	15,42	15,35
5	14,46	15,29	15,33	15,07
6	20,35	19,24	19,23	19,37
7	19,11	19,71	19,73	19,78
8	16,88	17,91	17,91	17,93
9	17,2	18,03	18,04	18,02
10	15,8	16,63	16,64	16,4
11	16,77	17,43	17,45	17,53
12	13,72	14,12	14,12	14,23
13	13,24	13,61	13,62	13,74
14	17,33	17,36	17,39	17,41
15	17,32	17,61	17,63	17,68
16	17,24	17,9	17,91	17,84
17	15,83	16,41	16,42	16,21
18	17,82	17,4	17,45	17,73
19	20,32	20,17	20,2	20,42
20	18,05	18,29	18,29	18,48
21	18,66	18,97	18,97	19,26
22	18,66	19,66	19,6	19,44
23	17,65	17,6	17,61	17,6
Erro médio		-0,38	-0,39	-0,41
$\sqrt{EQM}$		0,62	0,62	0,58



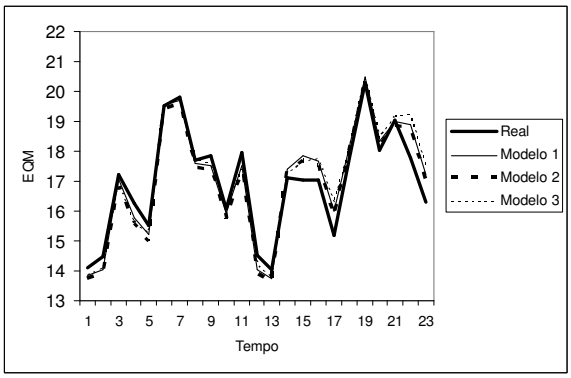
Quadro 5.7. Valores observados e preditos da armazenagem de água por tempo para o ponto 7.

Ponto 7				
Tempo	Real	Modelo 1	Modelo 2	Modelo 3
1	14,08	13,92	13,98	14,02
2	14,05	14,05	14,12	14,2
3	17,66	16,72	16,93	16,99
4	15,88	15,69	15,81	15,62
5	15,31	15,12	15,35	15,36
6	19,52	19,78	19,95	19,69
7	19,67	19,76	19,8	19,96
8	17,86	17,82	17,9	17,99
9	17,75	17,75	17,85	17,91
10	16,73	16,01	16,17	16,29
11	17,8	17,59	17,65	17,65
12	14,37	14,14	14,24	14,43
13	13,75	13,8	13,87	14,02
14	15,5	18,24	18,2	17,56
15	16,82	17,83	17,92	17,91
16	17,1	17,85	17,92	18,03
17	15,78	16,28	16,37	16,55
18	17,28	18,28	18,28	18,27
19	20,18	20,71	20,67	20,76
20	17,66	18,44	18,46	18,68
21	18,76	19,07	19,09	19,38
22	18,75	19,17	19,14	19,47
23	16,9	17,27	17,3	17,64
Erro médio		-0,27	-0,34	-0,40
$\sqrt{\text{EQM}}$		0,76	0,75	0,71



Quadro 5.8. Valores observados e preditos da armazenagem de água por tempo para o ponto 8.

Ponto 8				
Tempo	Real	Modelo 1	Modelo 2	Modelo 3
1	14,1	13,86	13,73	13,87
2	14,47	14,03	13,92	14,09
3	17,22	17,06	16,83	16,93
4	16,25	15,77	15,58	15,55
5	15,53	15,22	15,01	15,27
6	19,51	19,57	19,4	19,55
7	19,83	19,72	19,59	19,75
8	17,7	17,61	17,45	17,71
9	17,86	17,53	17,39	17,56
10	16,09	15,93	15,75	15,97
11	17,95	17,53	17,35	17,41
12	14,53	14,05	13,91	14,19
13	14,04	13,75	13,62	13,76
14	17,11	17,4	17,32	17,22
15	17,02	17,87	17,71	17,64
16	17,02	17,68	17,51	17,73
17	15,18	15,99	15,8	16,27
18	17,77	18,13	17,99	18,02
19	20,3	20,48	20,31	20,48
20	18,05	18,33	18,18	18,44
21	19,03	19	18,86	19,16
22	17,76	18,89	18,74	19,23
23	16,3	17,21	16,93	17,43
Erro médio		-0,09	0,08	-0,11
$\sqrt{\text{EQM}}$		0,49	0,48	0,56



Neste exemplo de aplicação podemos concluir que os erros de predição associados aos Modelos 1, 2 e 3 são menores nos locais onde a característica de interesse no ponto de predição e na vizinhança correspondente têm um comportamento semelhante. Aparentemente não existe diferença significativa entre os Modelos 1, 2 e 3.

A Tabela 5.4 mostra as principais medidas estatísticas para descrever os erros globais associados aos três modelos. Os erros foram calculados pela diferença entre os valores

observados e os valores preditos da armazenagem de água nos 23 instantes de tempo e para os 8 pontos amostrais de validação.

Tabela 5.4 – Análise descritiva dos erros.

Método	Média	Média absoluta	Desvio padrão	Mínimo	Máximo
Høst et al. (1995): EQM – Modelo 1	-0,344	0,581	0,757	-3,108	1,578
Høst et al. (1995): *di* – Modelo 2	-0,357	0,609	0,609	-3,172	1,572
Gneiting (2002): Separável – Modelo 3	-0,355	0,594	0,594	-3,183	1,787

Os resultados mostram que os erros médios de predição são semelhantes entre os três modelos.

A próxima seção apresenta o Caso 2 de análise dos dados.

5.4 Análise: Caso 2

No Caso 2 de análise o objetivo é predizer a armazenagem de água nos quatro últimos tempos (20, 21, 22 e 23) para os 8 pontos de validação escolhidos aleatoriamente na base de dados original. O banco de dados usado no ajuste dos modelos tem informação da variável nos 40 pontos amostrais nos instantes de tempo de 1 a 19. A análise descritiva usando estes dados é muito similar àquela discutida na seção 5.3 e os resultados não serão mostrados no texto.

As funções de covariância espaço-temporal separável e não-separável da família de Gneiting (2002) e o modelo baseado na proposta de Niu et al. (2003) definido na equação (3.44) foram ajustados aos dados e os resultados das predições são apresentados nas seções subseqüentes.

5.4.1 Ajuste pelo Modelo Baseado na Proposta de Niu et al. (2003)

Nesta seção mostramos o ajuste dos dados pelo modelo baseado na proposta de Niu et al. (2003) e dado pela equação (3.44). A quantidade de vizinhos do modelo foi determinada pelos critérios do EQM e da distância *di*. Pelo EQM obtemos uma vizinhança formada por 34 observações e este valor foi determinado pelo menor valor do EQM obtido usando as predições no tempo 20 nas 8 localizações de validação como mostra a Figura 5.13.

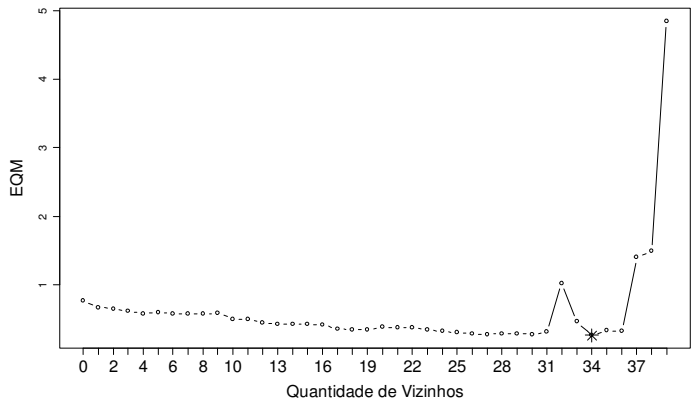


Figura 5.13. EQM de acordo com o número de vizinhos.

O modelo ajustado aos dados baseado no EQM é dado em (5.3):

$$\begin{aligned} Z(s,t) = \; & 9{,}58 + 0{,}58 \times Z(s,t-1) + 0{,}03 \times Z(v_1,t-1) - 0{,}02 \times Z(v_2,t-1) \\ & + 0{,}05 \times Z(v_3,t-1) - 0{,}03 \times Z(v_4,t-1) - 0{,}05 \times Z(v_5,t-1) + \ldots + \\ & - 0{,}07 \times Z(v_{33},t-1) - 0{,}21 \times Z(v_{34},t-1), \quad t = 20, 21, 22, 23 \end{aligned} \tag{5.3}$$

Observamos pelo modelo (5.3) que a informação de alguns vizinhos não parece contribuir significativamente na predição das novas observações e que o peso associado ao valor da característica de interesse no próprio local de predição no tempo anterior ao da predição é alto e igual a 0,58. A soma dos pesos associados aos vizinhos na predição é igual a -0,1373.

A Tabela 5.5 mostra o erro quadrático médio (EQM), a média dos resíduos (RES) e a média da soma de quadrados do erro (MSQE) pelo ajuste do modelo dado na equação (5.3) aos dados. Os erros são obtidos pela predição da armazenagem de água nas 8 localidades de validação em cada um dos tempos: 20, 21, 22 e 23.

Tabela 5.5 – Resultados da predição pelo ajuste do modelo dado em (5.3).

Tempo	EQM	RES	MSQE
20	0,26	-0,41	3,19
21	2,06	1,34	----
22	3,28	1,51	----
23	1,41	-0,20	----

A quantidade de vizinhos calculada pelo critério da distância *di* é igual a 11. Este valor corresponde ao número médio de localizações que estão a uma distância menor ou igual a 16,35 m em cada um dos pontos de predição. Este valor de *di* foi obtido pelo ajuste do

variograma circular ajustado aos dados no tempo 19 (ver equação 5.4 e Figura 5.14) e é igual ao parâmetro de alcance estimado.

$$\gamma(h) = \begin{cases} 0{,}2652^2 + 0{,}3941(\Gamma(h)) & , h > 0 \\ 0 & , h \leq 0 \end{cases} \tag{5.4}$$

onde $\Gamma(h) = \dfrac{2\left\{\left(\theta\sqrt{1-\theta^2}\right) + sen^{-1}\sqrt{\theta}\right\}}{\pi}$ e $\theta = \min\left(\dfrac{h}{16{,}35}, 1\right)$.

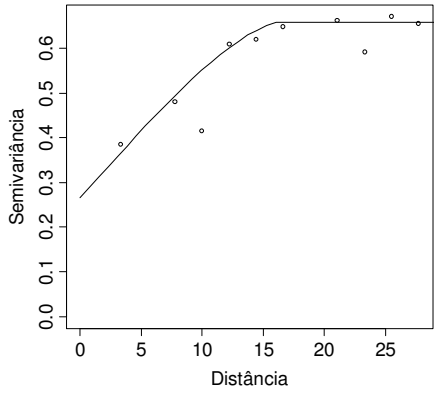


Figura 5.14. Variograma teórico ajustado aos dados no tempo 19.

A Figura 5.15 mostra o EQM de acordo com a distância. Observamos que a distância *di* igual a 16,35 m é uma escolha razoável. O EQM é calculado usando as predições nas 8 localizações de validação no tempo 20 (primeiro tempo de predição).

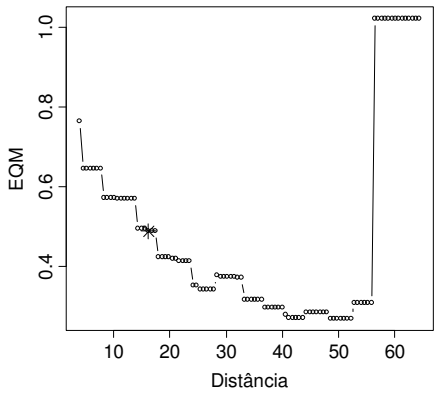


Figura 5.15. EQM de acordo com a distância *di*.

O modelo baseado nas idéias de Niu et al. (2003) ajustado aos dados pelo critério da distância *di* é dado em (5.5):

$$\begin{aligned} Z(s,t) = \; & 9{,}64 + 0{,}63 \times Z(s,t-1) + 0{,}05 \times Z(v_1,t-1) + 0{,}001 \times Z(v_2,t-1) \\ & - 0{,}02 \times Z(v_3,t-1) - 0{,}04 \times Z(v_4,t-1) - 0{,}06 \times Z(v_5,t-1) + \ldots + \\ & - 0{,}10 \times Z(v_{10},t-1) - 0{,}01 \times Z(v_{11},t-1), \quad t = 20,21,22,23 \end{aligned} \tag{5.5}$$

Pelo modelo (5.5) observamos que o peso dos vizinhos na predição temporal das observações é pequeno. A soma dos pesos correspondentes aos vizinhos é igual a -0,1943.

A Tabela 5.6 apresenta os resultados do ajuste pelo modelo (5.5) baseado na distância *di* de acordo com o tempo.

Tabela 5.6 – Resultados da predição pelo ajuste do modelo dado em (5.5).

Tempo	EQM	RES	MSQE
20	0,49	-0,63	3,29
21	1,22	0,96	----
22	2,05	1,12	----
23	1,52	-0,52	----

Os modelos dados pelas equações (5.3) e (5.5) serão denotados respectivamente por: Modelo 4 e Modelo 5.

Observamos pelas Tabelas 5.4 e 5.5 que os resultados pelo ajuste dos Modelos 4 e 5 são similares, porém o número de parâmetros do Modelo 5 é menor que do Modelo 4. O Modelo 5 utiliza a informação somente dos 11 vizinhos mais próximos, enquanto que no Modelo 4 a predição temporal é calculada usando a informação de 34 vizinhos.

A seção seguinte apresenta o ajuste dos dados pelas funções de covariância da família de Gneiting (2002).

5.4.2 Ajuste por Funções de Covariância da Família de Gneiting (2002)

O modelo separável mostrado na equação (4.2) estimado pelo método de máxima verossimilhança usando os dados nas 40 localizações nos tempos de 1 a 19 é dado por (5.6):

$$C(h,u) = \frac{46{,}95}{\left(0{,}7560|u|^{1{,}5758}+1\right)^{0{,}0633}} \exp\left(-0{,}0013\|h\|^{0{,}9263}\right) + \frac{0{,}2484}{\left(0{,}7560|u|^{1{,}5758}+1\right)^{0{,}0633}} \tag{5.6}$$

O modelo não-separável mostrado na equação (4.3) ajustado aos dados se reduz ao modelo separável dado em (5.6), pois o parâmetro β estimado é igual a zero.

As funções de verossimilhança condicional e perfilhada são mostradas na Figura 5.16. Observamos que o máximo para ambas as funções ocorre quando $\beta = 0$.

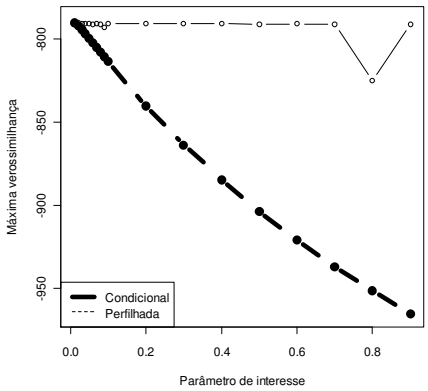


Figura 5.16. Máxima verossimilhança condicional e perfilhada.

O modelo separável (5.6) será denotado por Modelo 6. A comparação dos Modelos 4, 5 e 6 é feita na próxima seção.

5.4.3 Comparação dos Modelos Ajustados: Modelos 4, 5 e 6

A Tabela 5.7 compara os ajustes dos dados pelos Modelos 4, 5 e 6. Os dois primeiros se referem ao ajuste pelo modelo baseado na proposta de Niu et al. (2003) e que utiliza respectivamente o critério do EQM e a distância *di* para o cálculo do número de vizinhos. O terceiro modelo corresponde ao ajuste pela função de covariância separável da família de Gneiting (2002) dada na equação (5.6). A tabela apresenta o erro quadrático médio (EQM) e a média dos resíduos (RES) calculados usando as predições nos oito pontos amostrais de validação nos instantes de tempo: 20, 21, 22 e 23.

Tabela 5.7 – Comparação dos modelos 4, 5 e 6.

Modelo	EQM	RES
Niu et al. (2003): EQM - Modelo 4	1,75	0,56
Niu et al. (2003): *di* - Modelo 5	1,32	0,23
Gneiting (2002): Separável - Modelo 6	1,43	-0,63

Aparentemente não existe diferença entre os modelos no que se refere às predições calculadas nos 8 pontos de validação para os instantes de tempo 20, 21, 22 e 23.

Os Quadros de 5.9 a 5.12 apresentam os valores observados (ou reais) e os valores preditos da armazenagem de água no solo pelos Modelos 4, 5 e 6 para as 8 localidades de validação conforme o tempo: 20, 21, 22 e 23. No tempo 20 (ver Quadro 5.9) os valores preditos pelos Modelos 4 e 5 nas 8 localidades de validação são semelhantes com os valores reais. A predição feita pelo Modelo 6 no tempo 20 superestima os valores verdadeiros da armazenagem de água em todas as localidades.

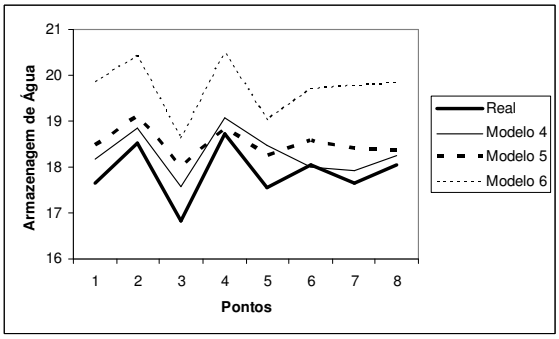
No tempo 21 (ver Quadro 5.10) as melhores predições correspondem ao Modelo 6. Os Modelos 4 e 5 subestimam os valores reais da característica de interesse.

Os valores preditos no tempo 22 (ver Quadro 5.11) considerando os três modelos não foram boas, com exceção da predição no ponto 3 onde o valor predito é próximo do valor observado.

As predições no tempo 23 (ver Quadro 5.12) são semelhantes entre os modelos e razoáveis para cada um dos pontos, exceto no ponto 3 no qual nenhum dos ajustes foi considerado adequado.

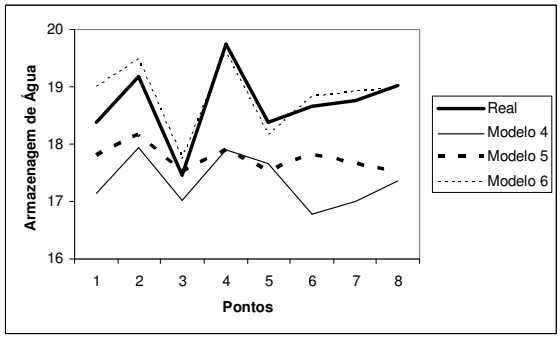
Quadro 5.9. Valores observados e preditos da armazenagem de água por ponto para o instante de tempo 20.

Tempo 20				
Ponto	Real	Modelo 4	Modelo 5	Modelo 6
1	17,64	18,18	18,47	19,86
2	18,51	18,84	19,13	20,43
3	16,84	17,57	18,02	18,61
4	18,72	19,08	18,86	20,51
5	17,54	18,48	18,24	19,03
6	18,05	18,01	18,58	19,70
7	17,66	17,93	18,40	19,78
8	18,05	18,24	18,36	19,83
Erro médio		-0,41	-0,63	-1,84
$\sqrt{\text{EQM}}$		0,51	0,70	1,86



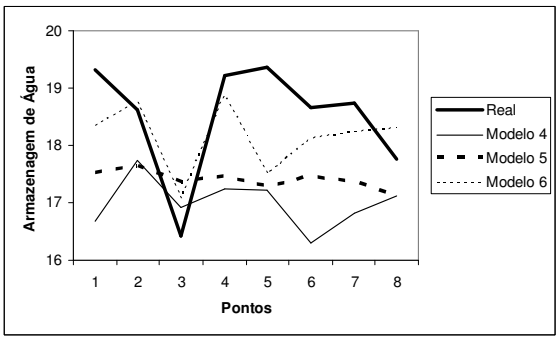
Quadro 5.10. Valores observados e preditos da armazenagem de água por ponto para o instante de tempo 21.

Tempo 21				
Ponto	Real	Modelo 4	Modelo 5	Modelo 6
1	18,38	17,15	17,80	19,01
2	19,17	17,94	18,15	19,48
3	17,47	17,02	17,53	17,74
4	19,73	17,91	17,89	19,59
5	18,37	17,65	17,54	18,17
6	18,66	16,79	17,81	18,82
7	18,76	17,01	17,65	18,91
8	19,03	17,36	17,50	18,97
Erro médio		1,34	0,96	-0,14
$\sqrt{\text{EQM}}$		1,43	1,11	0,29



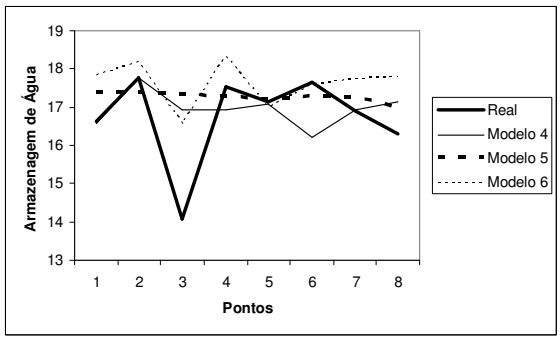
Quadro 5.11. Valores observados e preditos da armazenagem de água por ponto para o instante de tempo 22.

Tempo 22				
Ponto	Real	Modelo 4	Modelo 5	Modelo 6
1	19,33	16,69	17,51	18,35
2	18,63	17,73	17,65	18,75
3	16,42	16,92	17,37	17,07
4	19,23	17,25	17,47	18,88
5	19,35	17,23	17,28	17,50
6	18,66	16,30	17,45	18,12
7	18,75	16,82	17,35	18,24
8	17,76	17,12	17,13	18,31
Erro médio		1,51	1,12	0,37
$\sqrt{\text{EQM}}$		1,81	1,43	0,85



Quadro 5.12. Valores observados e preditos da armazenagem de água por ponto para o instante de tempo 23.

Tempo 23				
Ponto	Real	Modelo 4	Modelo 5	Modelo 6
1	16,62	16,58	17,38	17,84
2	17,76	17,77	17,38	18,20
3	14,07	16,94	17,33	16,57
4	17,52	16,94	17,28	18,34
5	17,14	17,07	17,19	16,99
6	17,65	16,21	17,28	17,60
7	16,90	16,92	17,25	17,73
8	16,30	17,13	17,00	17,80
Erro médio		-0,20	-0,52	-0,89
$\sqrt{\text{EQM}}$		1,19	1,23	1,20



A Tabela 5.8 mostra as medidas de tendência central e de dispersão para descrever os erros globais associados a cada um dos três Modelos: 4, 5 e 6. Os erros globais são calculados pela diferença entre os valores observados e os valores preditos nas 8 localidades de validação nos instantes de tempo 20, 21, 22 e 23.

Tabela 5.8 – Análise descritiva dos erros.

Método	Média	Média absoluta	Desvio padrão	Mínimo	Máximo
Niu et al. (2003): EQM – Modelo 4	0,559	1,033	1,218	-2,87	2,64
Niu et al. (2003): *di* – Modelo 5	0,233	0,931	1,142	-3,26	2,07
Gneiting (2002): Separável – Modelo 6	-0,627	0,929	1,032	-2,50	1,85

O menor erro médio corresponde ao ajuste feito pelo Modelo 5. Para alguns pontos de validação o Modelo 6 apresentou melhores resultados.

A seção seguinte trata do Caso 3 de análise.

5.5 Análise: Caso 3

O objetivo do Caso 3 de análise é predizer a armazenagem de água nas 8 localidades de validação nos instantes de tempo: 20, 21, 22 e 23. O banco de dados usado para testar a adequação dos modelos é formado pelos 32 pontos amostrais e para cada ponto tem-se a informação da característica de interesse no período de 1 a 19. Inicialmente foi realizada uma análise descritiva dos dados e os resultados foram muito similares com aqueles apresentados na seção 5.3.

A predição nos locais e tempos descritos no parágrafo precedente é calculada pelo ajuste de dois modelos distintos aos dados. O primeiro modelo combina as metodologias de geoestatística e de séries temporais em duas etapas como mostrado no Capítulo 3, e o segundo ajuste se refere às funções de covariância espaço-temporal separável e não-separável da família de Gneiting (2002). As próximas seções apresentam os resultados destes ajustes.

5.5.1 Combinação dos Modelos de Geoestatística e de Séries Temporais

Nesta seção mostramos o ajuste dos dados pela combinação dos modelos de geoestatística e de séries temporais. Primeiramente foi ajustado o modelo de Høst et al. (1995) aos dados para predizer as observações nas 8 localidades de validação nos instantes de tempo de 1 a 19. Em seguida o banco de dados foi atualizado com estas predições e o modelo baseado nas idéias de Niu et al. (2003) foi ajustado para predizer a armazenagem de água nestes mesmos 8 pontos amostrais de validação nos tempos 20, 21, 22 e 23.

Na metodologia de Høst et al. (1995) o variograma teórico ajustado à componente *F* foi o cúbico e os parâmetros estimados para a média, o efeito pepita, a variância e o alcance são respectivamente: 16,82; 0,27; 0,15 e 14,44, i.e.:

$$\gamma(h) = 0{,}27 + 0{,}15\begin{cases} 7\left(\frac{h}{14{,}44}\right)^2 - 8{,}75\left(\frac{h}{14{,}44}\right)^3 + 3{,}5\left(\frac{h}{14{,}44}\right)^5 - 0{,}75\left(\frac{h}{14{,}44}\right)^7, & \text{se } h < 14{,}44 \\ 0, & \text{caso contrário} \end{cases} \quad (5.7)$$

A Figura 5.17 mostra o variograma teórico ajustado à componente *F*.

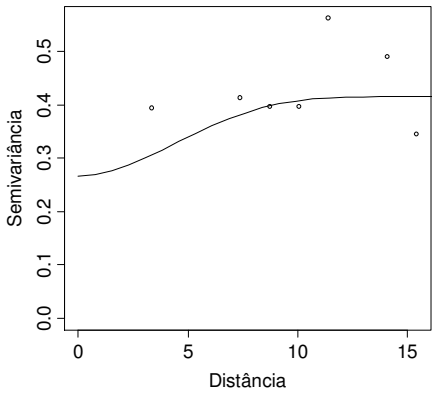


Figura 5.17. Variograma teórico ajustado à componente *F*.

A predição da armazenagem de água nas 8 localidades de validação nos instantes de tempo de 1 a 19 foi calculada de duas maneiras de acordo com o critério para a escolha do número de vizinhos. Pelo EQM o número de vizinhos é igual a 7 como mostra a Figura 5.18, e a quantidade de vizinhos baseada na distância *di* para os pontos amostrais de 1 a 8 é respectivamente: 11, 9, 10, 12, 10, 9, 9 e 10.

O valor de *di* é igual 14,44 m e corresponde ao parâmetro de alcance estimado pelo variograma cúbico ajustado à componente *F* (ver equação (5.7)). A Figura 5.19 apresenta o EQM de acordo com a distância e podemos observar que o valor de *di* igual a 14,44 m é uma escolha razoável, apesar de que distâncias próximas do intervalo entre 10 e 14 m resultam em menores erros de predição.

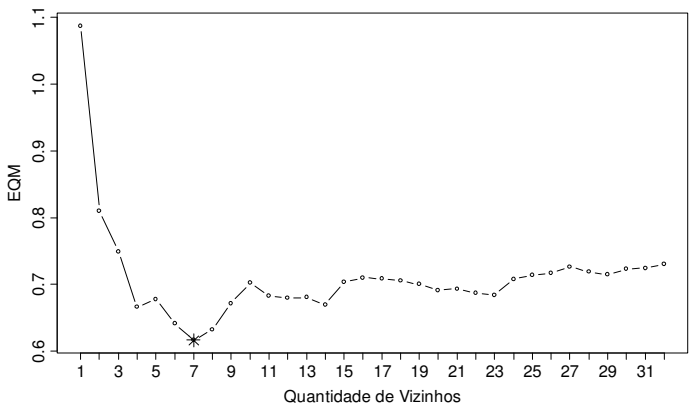


Figura 5.18. EQM de acordo com a quantidade de vizinhos.

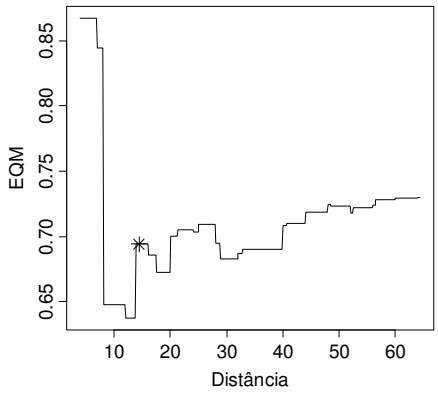


Figura 5.19. EQM de acordo com a distância.

O modelo de Høst et al. (1995) ajustado aos dados baseado no critério do EQM e na distância *di* são denotados respectivamente por: Modelo 7 e Modelo 8. A adequação destes

ajustes foi avaliada pela validação cruzada, i.e., um a um dos 32 pontos da base de dados foi retirado e a observação predita nos tempos de 1 a 19 usando os Modelos 7 e 8. A média dos resíduos pelo ajuste de ambos os modelos foi próxima de zero e a distribuição dos mesmos foi aproximadamente normal.

A Tabela 5.9 compara os ajustes feitos pelos Modelos 7 e 8 através do erro quadrático médio (EQM) e da média dos resíduos (RES) calculados usando as predições nas 8 localizações de validação nos tempos de 1 a 19.

Tabela 5.9 – Comparação dos modelos 7 e 8.

Modelo	EQM	RES
Høst et al. (1995): EQM – Modelo 7	0,6162	-0,2917
Høst et al. (1995): *di* – Modelo 8	0,6943	-0,3109

Para predizer a armazenagem de água nos 8 pontos amostrais de validação nos tempos 20, 21, 22 e 23 a base de dados foi atualizada com as predições calculadas pelo ajuste dos Modelos 7 e 8 e a metodologia de séries temporais foi aplicada aos dados.

Na segunda etapa de predição o modelo baseado nas idéias de Niu et al. (2003) e dado pela equação (3.44) foi ajustado a esta nova base de dados e a quantidade de vizinhos do modelo foi calculada de duas maneiras distintas. A primeira é baseada no menor valor do EQM obtido pelas predições no tempo 20 para as 8 localidades de validação, e a segunda forma considera o número médio de vizinhos que estão a uma distância menor ou igual a *di* em cada um dos pontos amostrais onde se deseja fazer a predição. A distância *di* é obtida pelo ajuste do variograma teórico aos dados no tempo 19 e corresponde ao parâmetro de alcance estimado.

O modelo temporal ajustado pelo critério do EQM é dado em (5.8):

$$\begin{aligned} Z(s,t) = \; & 9{,}79 + 0{,}56 \times Z(s,t-1) - 0{,}03 \times Z(v_1,t-1) - 0{,}05 \times Z(v_2,t-1) \\ & + 0{,}03 \times Z(v_3,t-1) - 0{,}07 \times Z(v_4,t-1) - 0{,}06 \times Z(v_5,t-1) + \ldots + \\ & - 0{,}43 \times Z(v_{35},t-1) - 0{,}22 \times Z(v_{36},t-1), \quad t = 20, 21, 22, 23 \end{aligned} \tag{5.8}$$

Observamos pela equação (5.8) que o peso associado a alguns vizinhos é pequeno e utiliza-se quase toda a vizinhança na predição (34 pontos). O valor da armazenagem de água no ponto de predição no tempo anterior ao da predição é importante, visto que o peso atribuído a esta informação é alto e igual a 0,56. A soma dos pesos associados aos vizinhos é igual a -0,12.

A Tabela 5.10 resume os resultados da predição aplicando o modelo dado em (5.8) aos dados. Os erros mostrados nesta tabela foram calculados utilizando as predições nas 8 localizações de validação de acordo com o tempo.

Tabela 5.10 – Resultados da predição pelo ajuste do modelo dado em (5.8).

Tempo	EQM	RES	MSQE
20	0,51	-0,37	3,09
21	3,26	1,57	----
22	4,02	1,80	----
23	1,67	-0,12	----

A escolha do número de vizinhos pelo critério da distância *di* é calculada utilizando o valor do parâmetro de alcance estimado pelo modelo de variograma teórico ajustado aos dados no instante de tempo 19. O variograma esférico foi ajustado a estes dados (ver Figura 5.20) e a equação do modelo pode ser escrita como em (5.9):

$$\gamma(h)=\begin{cases}0 & , \ h\le 0\\ 0{,}37^2+0{,}26\left[\dfrac{3h}{43{,}02}-\dfrac{1}{2}\left(\dfrac{h}{21{,}51}\right)^3\right] & , \ 0<h<21{,}51\\ 0{,}37^2+0{,}26 & , \ h\ge 21{,}51\end{cases} \tag{5.9}$$

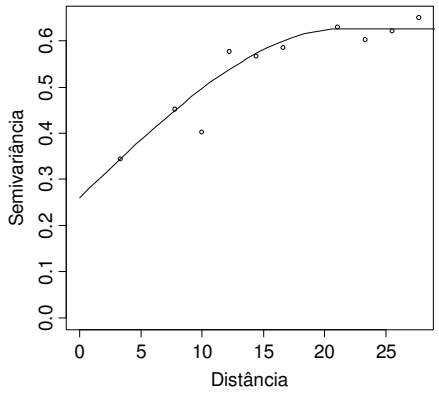


Figura 5.20. Variograma teórico ajustado aos dados no tempo 19.

O modelo ajustado aos dados usando o critério da distância *di* é dado em (5.10):

$$\begin{aligned}Z(s,t)= \ & 9{,}70+0{,}63\times Z(s,t-1)+0{,}02\times Z(v_1,t-1)+0{,}01\times Z(v_2,t-1)\\ & +0{,}04\times Z(v_3,t-1)-0{,}04\times Z(v_4,t-1)-0{,}03\times Z(v_5,t-1)+\ldots+\\ & -0{,}01\times Z(v_{15},t-1)-0{,}06\times Z(v_{16},t-1), \quad t=20,21,22,23\end{aligned} \tag{5.10}$$

Neste caso o número de vizinhos utilizados na predição é menor se comparado com o modelo (5.8) e igual a 16. A informação da característica de interesse no local de predição no tempo imediatamente anterior ao da predição é relevante e o peso atribuído a este dado é igual a 0,63. A soma dos pesos dos vizinhos é igual a -0,19.

A Figura 5.21 mostra o EQM de acordo com a distância. O ponto em destaque "*" corresponde à distância *di* igual a 21,51 m. Podemos observar que este valor resulta em um erro pequeno de predição se comparado com outras distâncias. É importante salientar que este gráfico é meramente ilustrativo não podendo ser construído em situações práticas (reais). O EQM é calculado usando as predições nas 8 localidades de validação no instante de tempo 20. Os resultados do ajuste pelo modelo dado na equação (5.10) são apresentados na Tabela 5.11.

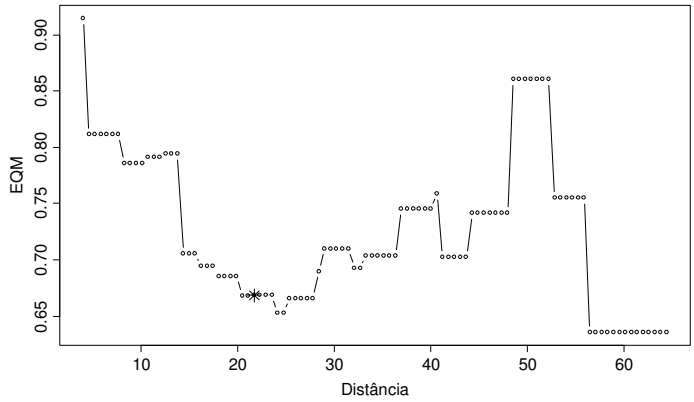


Figura 5.21. EQM de acordo com a distância.

Tabela 5.11 – Resultados da predição pelo ajuste do modelo dado em (5.10).

Tempo	EQM	RES	MSQE
20	0,67	-0,68	3,21
21	1,11	0,90	----
22	1,94	1,06	----
23	1,43	-0,57	----

Os modelos dados pelas equações (5.8) e (5.10) são denotados respectivamente por: Modelo 9 e Modelo 10.

Podemos inferir pela análise das Tabelas 5.9 e 5.10 que o ajuste pelo Modelo 10 é preferível ao ajuste pelo Modelo 9, pois este último resulta em erros maiores de predição além de utilizar uma quantidade maior de parâmetros na predição de novas observações.

A seção seguinte trata do ajuste dos dados pelas funções de covariância da família de Gneiting (2002). A comparação do ajuste por estas funções de covariância e pelos Modelos 9 e 10 são apresentados na seção 5.5.3.

5.5.2 Ajuste por Funções de Covariância da Família de Gneiting (2002)

A função de covariância espaço-temporal separável dada pela equação (4.2) foi ajustada a base de dados formada pelos 32 pontos amostrais nos instantes de tempo de 1 a 19. Os parâmetros foram estimados pelo método da máxima verossimilhança e a equação do modelo é dada em (5.11):

$$C(h,u)=\frac{47{,}60}{\left(0{,}7663|u|^{1{,}4838}+1\right)^{0{,}0583}}\exp\left(-0{,}0013\|h\|^{0{,}9398}\right)+\frac{0{,}2710}{\left(0{,}7663|u|^{1{,}4838}+1\right)^{0{,}0583}} \qquad (5.11)$$

A função de covariância não-separável mostrada na equação (4.3) também foi ajustada a estes dados, porém o parâmetro β estimado é igual a zero indicando a independência dos processos espacial e temporal. Desta forma o modelo não-separável se reduz ao modelo separável dado pela equação (5.11) e denotaremos este ajuste para o cálculo das predições por Modelo 11.

O estudo das verossimilhanças condicional e da perfilhada é mostrado na Figura 5.22.

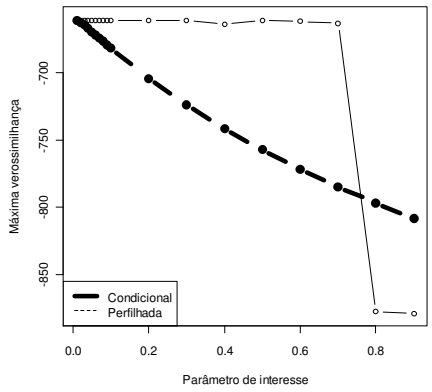


Figura 5.22. Máxima verossimilhança condicional e perfilhada.

A seção seguinte compara os ajustes aos dados feitos pelos Modelos 9, 10 e 11.

5.5.3 Comparação dos Modelos Ajustados: Modelos 9, 10 e 11

Nesta seção os Modelos 9, 10 e 11 são comparados utilizando os erros associados às predições da armazenagem de água nos 8 pontos amostrais de validação nos instantes de tempo 20, 21, 22 e 23. Os dois primeiros modelos se referem ao ajuste que combina as metodologias de geoestatística e de séries temporais baseado nos critérios respectivamente do

EQM e da distância *di* para o cálculo do número de vizinhos. O Modelo 11 é relativo ao ajuste dos dados pela função de covariância separável dada pela equação (5.11). A Tabela 5.12 resume os erros associados ao ajuste por cada um dos três modelos.

Tabela 5.12 – Comparação dos modelos 9, 10 e 11.

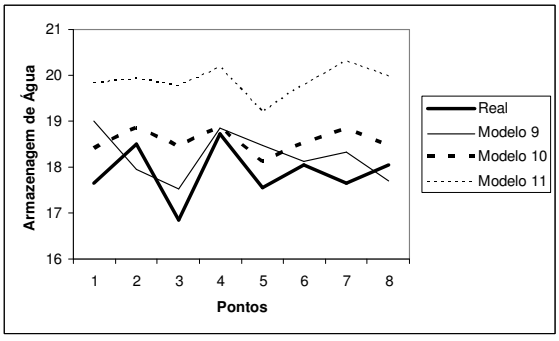
Modelo	EQM	RES
Combinado: EQM - Modelo 9	2,36	0,72
Combinado: *di* - Modelo 10	1,29	0,18
Gneiting (2002): Separável - Modelo 11	2,11	-0,86

Observamos pela Tabela 5.12 que o melhor ajuste corresponde ao Modelo 10. Aparentemente não existe diferença entre os Modelos 9 e 11.

Os Quadros de 5.13 a 5.16 mostram os valores observados e os valores preditos da armazenagem de água para cada uma das 8 localizações de validação de acordo com o instante de tempo: 20, 21, 22 e 23. As conclusões são similares com aquelas apresentadas na seção 5.4.3 referentes aos Quadros de 5.9 a 5.12.

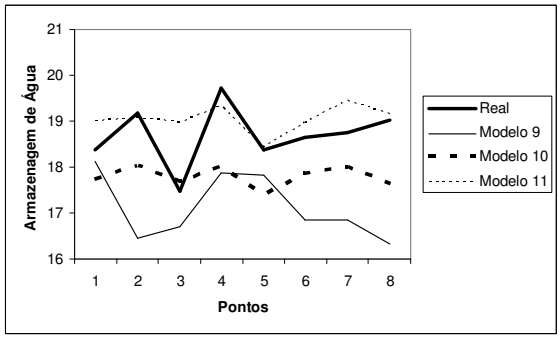
Quadro 5.13. Valores observados e preditos da armazenagem de água por ponto para o instante de tempo 20.

Tempo 20				
Ponto	Real	Modelo 4	Modelo 5	Modelo 6
1	17,64	19,01	18,40	19,82
2	18,51	17,96	18,85	19,94
3	16,84	17,53	18,44	19,78
4	18,72	18,84	18,89	20,18
5	17,54	18,47	18,12	19,21
6	18,05	18,14	18,52	19,79
7	17,66	18,32	18,82	20,30
8	18,05	17,71	18,44	19,98
Erro médio		-0,37	-0,68	-2,00
$\sqrt{EQM}$		0,72	0,82	2,06



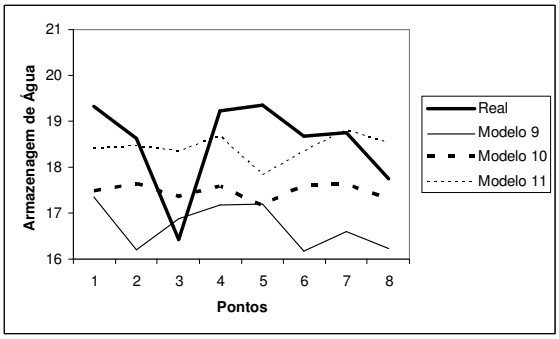
Quadro 5.14. Valores observados e preditos da armazenagem de água por ponto para o instante de tempo 21.

Tempo 21				
Ponto	Real	Modelo 4	Modelo 5	Modelo 6
1	18,38	18,14	17,72	19,01
2	19,17	16,45	18,04	19,08
3	17,47	16,71	17,69	18,97
4	19,73	17,87	18,00	19,33
5	18,37	17,83	17,41	18,41
6	18,66	16,85	17,86	18,97
7	18,76	16,86	18,01	19,45
8	19,03	16,32	17,64	19,14
Erro médio		1,57	0,90	-0,35
$\sqrt{EQM}$		1,81	1,05	0,65



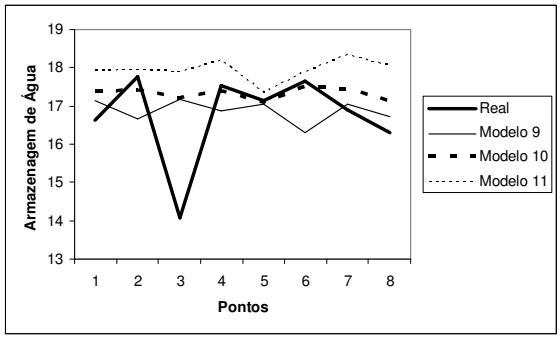
Quadro 5.15. Valores observados e preditos da armazenagem de água por ponto para o instante de tempo 22.

Tempo 22				
Ponto	Real	Modelo 4	Modelo 5	Modelo 6
1	19,33	17,34	17,46	18,40
2	18,63	16,19	17,64	18,44
3	16,42	16,86	17,35	18,36
4	19,23	17,18	17,59	18,69
5	19,35	17,21	17,15	17,81
6	18,66	16,16	17,60	18,35
7	18,75	16,59	17,62	18,81
8	17,76	16,22	17,28	18,51
Erro médio		1,80	1,06	0,10
$\sqrt{EQM}$		2,00	1,39	1,00



Quadro 5.16. Valores observados e preditos da armazenagem de água por ponto para o instante de tempo 23.

Tempo 23				
Ponto	Real	Modelo 4	Modelo 5	Modelo 6
1	16,62	17,13	17,38	17,93
2	17,76	16,66	17,42	17,95
3	14,07	17,16	17,20	17,89
4	17,52	16,88	17,39	18,20
5	17,14	17,04	17,08	17,36
6	17,65	16,31	17,50	17,89
7	16,90	17,04	17,43	18,33
8	16,30	16,71	17,12	18,04
Erro médio		-0,12	-0,57	-1,20
$\sqrt{EQM}$		1,29	1,20	1,66



A Tabela 5.13 mostra as medidas de tendência central e de dispersão para descrever os erros globais associados a cada um dos três Modelos: 9, 10 e 11. Os erros são calculados usando as predições nas 8 localidades de validação nos instantes de tempo 20, 21, 22 e 23.

Tabela 5.13 – Análise descritiva dos erros.

Método	Média	Média absoluta	Desvio padrão	Mínimo	Máximo
Combinação: EQM – Modelo 9	0,718	1,246	1,382	-3,09	2,72
Combinação: *di* – Modelo 10	0,175	0,917	1,138	-3,13	2,20
Gneiting (2002): Separável – Modelo 11	-0,864	1,114	1,186	-3,82	1,54

O Modelo 5 é o que apresenta os melhores ajustes de acordo com os erros de predição se comparados com os Modelos 4 e 6. A variabilidade dos erros é grande para os três modelos considerados na análise.

Pelas análises apresentadas neste capítulo observamos que não existe diferença entre os modelos de Høst et al. (1995) e a função de covariância separável proposta por Gneiting (2002) na interpolação espacial da armazenagem de água no solo com citros em tempos

amostrados e os erros obtidos ajustando esses modelos são pequenos. O comportamento dos vizinhos no local de predição têm influência na qualidade do ajuste desses modelos como observamos na predição da armazenagem de água no ponto 3 de validação.

As predições em tempos futuros em localizações amostradas, e as predições em pontos e tempos não observados na amostra foram razoáveis conforme os modelos ajustados nesse capítulo. O modelo baseado nas idéias de Niu et al. (2003) e a combinação dos modelos de geoestatística e de séries temporais, ambos utilizando o EQM no cálculo do número de vizinhos, forneceram erros menores na predição das observações se comparado com o ajuste pela função de covariância separável da classe de Gneiting (2002).

Capítulo 6

Estudo de Caso: Incidência de AIDS no Estado de Minas Gerais

Neste capítulo aplicamos os modelos apresentados no Capítulo 3 aos dados de incidência de AIDS no estado de Minas Gerais (MG).

6.1 Introdução

Os dados de incidência de AIDS foram analisados de três formas distintas: Casos 1, 2 e 3 ajustando-se o modelo proposto por Høst et al. (1995) com as modificações apresentadas previamente, o modelo baseado nas idéias de Niu et al. (2003), a combinação destes modelos e as funções de covariância separável e não-separável da família de Gneiting (2002). As seções seguintes mostram os resultados destes ajustes, bem como a descrição dos dados.

6.2 Descrição dos Dados

O estado de Minas Gerais é dividido geograficamente (mas não politicamente) em sessenta e seis microrregiões[14]: Aimorés, Alfenas, Almenara, Andrelândia, Araçuaí, Araxá, Barbacena, Belo Horizonte, Bocaiúva, Bom Despacho, Campo Belo, Capelinha, Caratinga, Cataguases, Conceição do Mato Dentro, Conselheiro Lafaiete, Curvelo, Diamantina, Divinópolis, Formiga, Frutal, Guanhães, Governador Valadares, Grão Mogol, Ipatinga, Itabira, Itaguara, Itajubá, Ituiutaba, Janaúba, Januária, Juiz de Fora, Lavras, Manhuaçu, Mantena, Montes Claros, Muriaé, Nanuque, Oliveira, Ouro Preto, Pará de Minas, Paracatu, Passos, Patos de Minas, Patrocínio, Peçanha, Pedra Azul, Pirapora, Piumhi, Poços de Caldas, Ponte Nova, Pouso Alegre, Salinas, Santa Rita do Sapucaí, São João Del-Rei, São Lourenço, São Sebastião do Paraíso, Sete Lagoas, Teófilo Otoni, Três Marias, Ubá, Uberaba, Uberlândia, Unaí, Varginha, Viçosa. A Figura 6.1 apresenta as microrregiões do estado de Minas Gerais.

[14] http://www.geominas.mg.gov.br/ (acesso em 01/2008).

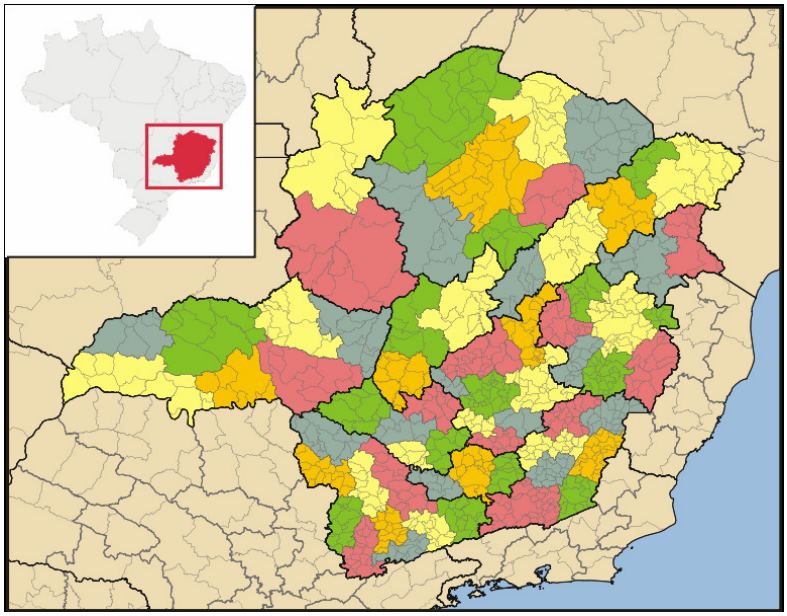


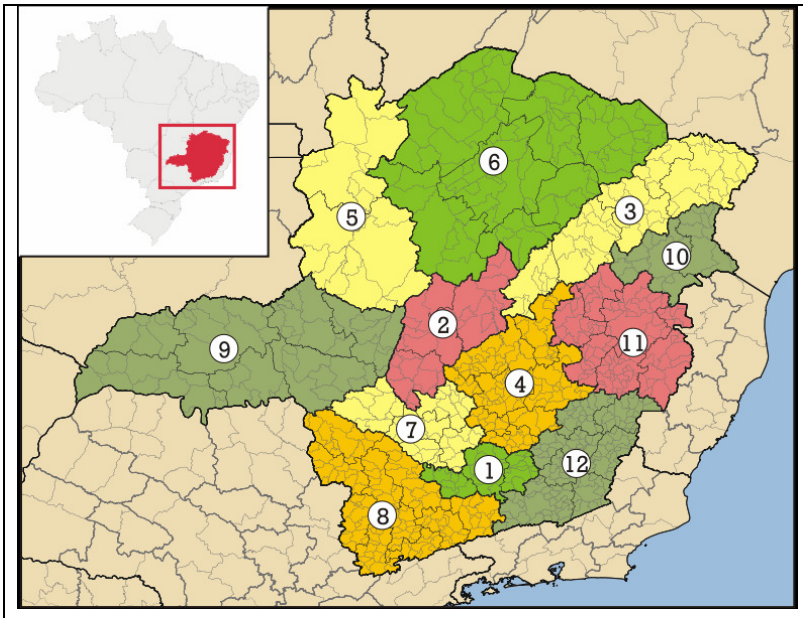
Figura 6.1. Microrregiões do estado de Minas Gerais.

Fonte: http://www.wikipedia.org/

Os dados, obtidos no site do DATASUS, se referem à quantidade de casos de AIDS no período entre 1991 e 2006 nas microrregiões do estado de Minas Gerais. Somente o município de Piumhi não apresenta informações sobre a incidência da doença. Assim, obtemos 65 posições no espaço e 16 períodos de tempo distintos. Os municípios são localizados pelas coordenadas de latitude e longitude. Na modelagem dos dados 20% (ou em valor absoluto, 13) dos municípios foram separados para validação, ou seja, para a verificação posterior da qualidade de ajuste dos modelos.

Os municípios separados para validação foram escolhidos aleatoriamente dentro de cada mesorregião do estado de Minas Gerais. De acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE), o estado de Minas Gerais é dividido em 12 mesorregiões: Campo das Vertentes, Central Mineira, Jequitinhonha, Metropolitana de Belo Horizonte, Noroeste de Minas, Norte de Minas, Oeste de Minas, Sul e Sudoeste de Minas, Triângulo Mineiro e Alto Paranaíba, Vale do Mucuri, Vale do Rio Doce, Zona da Mata. As mesorregiões de MG são apresentadas na Figura 6.2.

Desta forma escolhemos aleatoriamente um município dentro de cada mesorregião, exceto na mesorregião Metropolitana de Belo Horizonte, onde selecionamos 2 municípios para satisfazer a porcentagem de exclusão para validação que é igual a 20%. Os 13 municípios separados da análise são: Belo Horizonte, Campo Belo, Curvelo, Diamantina, Ipatinga, Itaguara, Lavras, Montes Claros, Patrocínio, Ponte Nova, Pouso Alegre, Teófilo Otoni, Unaí. A Figura 6.3 mostra as localizações desconsideradas nas análises para validação.



1. Campo das Vertentes
2. Central Mineira
3. Jequitinhonha
4. Metropolitana de Belo Horizonte
5. Noroeste de Minas
6. Norte de Minas
7. Oeste de Minas
8. Sul e Sudoeste de Minas
9. Triângulo Mineiro e Alto Paranaíba
10. Vale do Mucuri
11. Vale do Rio Doce
12. Zona da Mata

Figura 6.2. Mesorregiões do estado de Minas Gerais.

Fonte: http://www.wikipedia.org/

Figura 6.3. Municípios separados da análise para validação dos modelos.

A Tabela 6.1 mostra a média, o desvio-padrão, a soma do número de casos de AIDS nas microrregiões de MG no período de 1991 a 2006 e a respectiva população no ano de 2006.

Tabela 6.1 – Análise Descritiva dos Dados de Incidência de AIDS em MG.

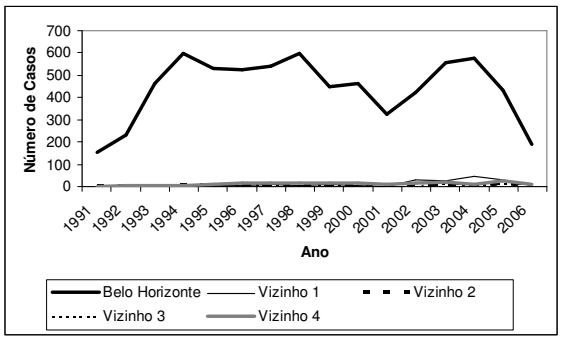
Município	Média	Desvio-padrão	Soma	População	Município	Média	Desvio-padrão	Soma	População
Aimorés	0,38	0,62	6	24.120	Manhuaçu	3	2,85	48	73.516
Alfenas	5,31	5,11	85	77.495	Mantena	1,69	3,03	27	25.105
Almenara	0,44	0,63	7	36.639	Montes Claros	11,56	7,06	185	348.990
Andrelândia	0,25	0,58	4	12.173	Muriaé	6,56	5,96	105	100.068
Araçuaí	0,75	1	12	37.109	Nanuque	1,69	1,66	27	40.530
Araxá	9,44	5,19	151	85.712	Oliveira	1,25	1,06	20	40.967
Barbacena	3,69	2,85	59	124.603	Ouro Preto	3	2,94	48	69.056
Belo Horizonte	440,44	143,96	7047	2.399.920	Paracatu	2,5	2,61	40	84.411
Bocaiúva	1,12	1,02	18	45.349	Pará de Minas	3,88	3,67	62	81.738
Bom Despacho	1,25	1,77	20	43.353	Passos	12,88	15,38	206	106.516
Campo Belo	1,81	2,01	29	52.633	Patos de Minas	9,19	6,05	147	139.357
Capelinha	0,44	0,73	7	35.420	Patrocínio	1,81	2,23	29	82.279
Caratinga	4,06	2,26	65	82.633	Peçanha	0,25	0,58	4	17.060
Cataguases	4,19	2,1	67	68.300	Pedra Azul	0,69	0,7	11	24.745
Conceição do Mato Dentro	0,12	0,34	2	18.579	Pirapora	2,81	1,83	45	53.219
Conselheiro Lafaiete	4,25	2,65	68	113.019	Poços de Caldas	24,5	22,06	392	154.474
Curvelo	3,31	2,63	53	73.791	Ponte Nova	3,75	3,77	60	57.344
Diamantina	1,75	1,57	28	44.223	Pouso Alegre	15,38	11,27	246	125.206
Divinópolis	13,44	10,78	215	207.981	Salinas	0,44	0,73	7	37.956
Formiga	3,69	2,96	59	67.174	Santa Rita do Sapucaí	1,94	1,48	31	34.920
Frutal	12,06	5,05	193	50.366	São João Del Rei	7,56	5,57	121	82.952
Governador Valadares	27,31	15,38	437	259.407	São Lourenço	7,12	4,38	114	42.145
Grão Mogol	0,19	0,4	3	15.625	São Sebastião do Paraíso	8,12	6,33	130	65.197
Guanhães	0,56	0,81	9	29.788	Sete Lagoas	12,44	13,16	199	215.068
Ipatinga	19,69	11,75	315	236.463	Teófilo Otoni	8,38	6,4	134	127.530
Itabira	11,44	6,56	183	107.721	Três Marias	0,69	1,01	11	25.170
Itaguara	0,69	1,14	11	11.767	Ubá	3,81	3,66	61	98.776
Itajubá	14,81	7,8	237	90.815	Uberaba	70,44	26,65	1127	285.093
Ituiutaba	13,25	5,62	212	92.428	Uberlândia	87,69	50,93	1403	600.367
Janaúba	0,88	1,26	14	70.093	Unaí	1,12	1,02	18	76.244
Januária	1,56	1,67	25	62.516	Varginha	7,44	3,41	119	124.501
Juiz de Fora	110,5	65,34	1768	509.126	Viçosa	2,19	1,8	35	74.606
Lavras	6,31	3,5	101	88.290					

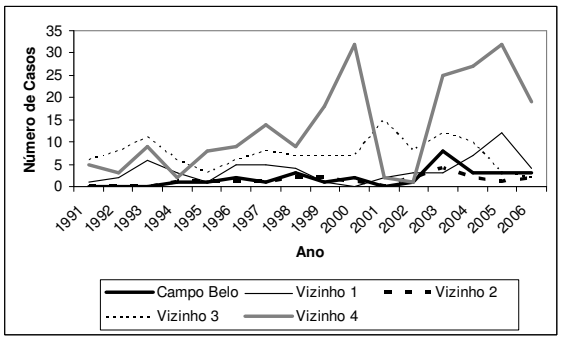
Os municípios com a incidência média de AIDS mais alta no período de 1991 a 2006 correspondem a: Belo Horizonte, Juiz de Fora, Uberlândia e Uberaba. Estes municípios também são os mais populosos com base no ano de 2006 se comparados com os outros municípios que compõem as microrregiões do estado de MG.

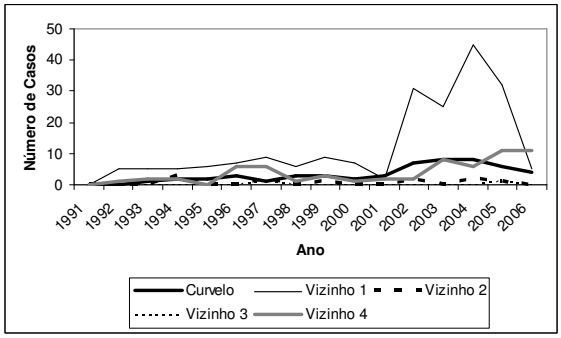
A menor quantidade de casos da doença está associada aos municípios de: Conceição do Mato Dentro, Grão Mogol, Andrelândia, Peçanha e Aimorés.

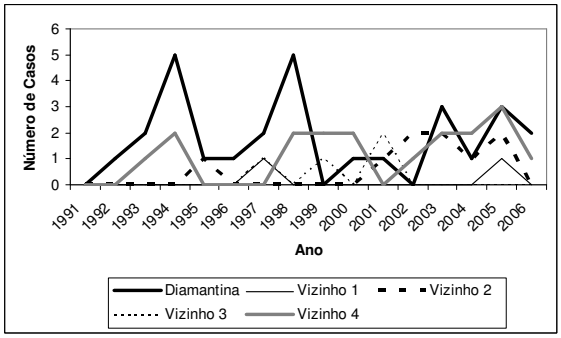
A variação no número de casos de AIDS entre os municípios é grande.

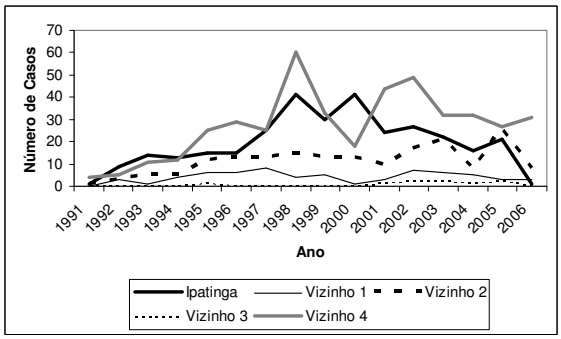
A Figura 6.4 mostra a incidência da doença nos 13 municípios desconsiderados da análise para validação e nos quatro vizinhos mais próximos ao longo do tempo.

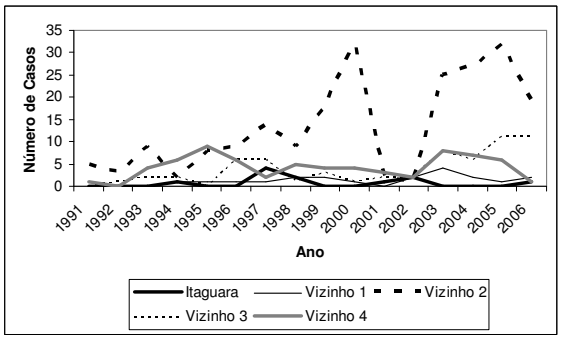

Número de Casos
Ano
Belo Horizonte
Vizinho 1
Vizinho 2
Vizinho 3
Vizinho 4

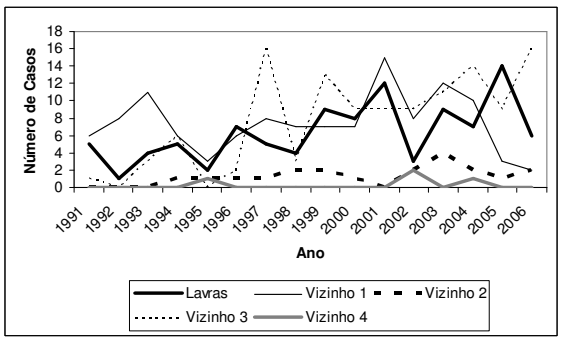

Número de Casos
Ano
Campo Belo
Vizinho 1
Vizinho 2
Vizinho 3
Vizinho 4

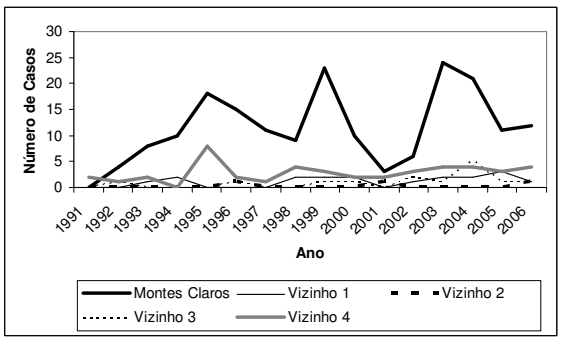

Número de Casos
Ano
Curvelo
Vizinho 1
Vizinho 2
Vizinho 3
Vizinho 4

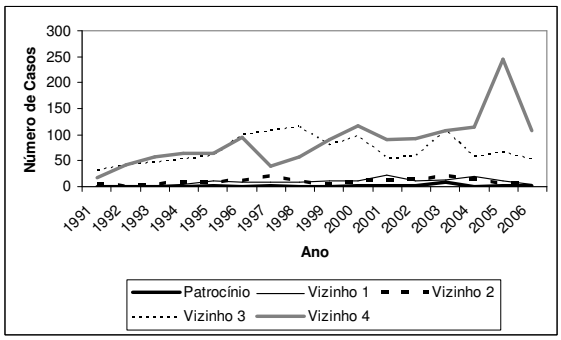

Número de Casos
Ano
Diamantina
Vizinho 1
Vizinho 2
Vizinho 3
Vizinho 4

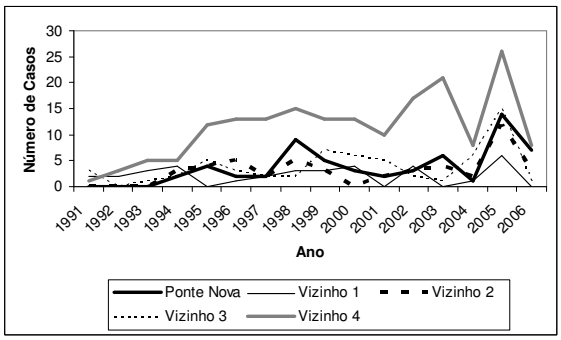

Número de Casos
Ano
Ipatinga
Vizinho 1
Vizinho 2
Vizinho 3
Vizinho 4


Número de Casos
Ano
Itaguara
Vizinho 1
Vizinho 2
Vizinho 3
Vizinho 4


Número de Casos
Ano
Lavras
Vizinho 1
Vizinho 2
Vizinho 3
Vizinho 4


Número de Casos
Ano
Montes Claros
Vizinho 1
Vizinho 2
Vizinho 3
Vizinho 4


Número de Casos
Ano
Patrocínio
Vizinho 1
Vizinho 2
Vizinho 3
Vizinho 4


Número de Casos
Ano
Ponte Nova
Vizinho 1
Vizinho 2
Vizinho 3
Vizinho 4

Continua

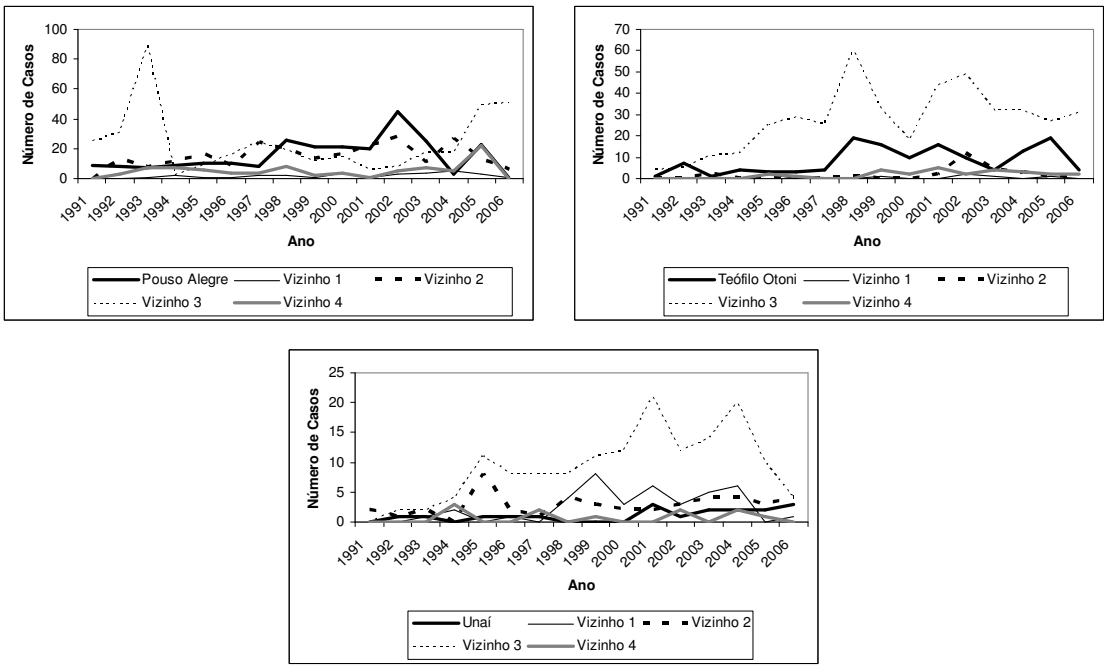


Figura 6.4. Número de casos de AIDS nos 13 municípios de validação no período de 1991 a 2006 e nos 4 vizinhos mais próximos.

Os dados foram analisados nesta dissertação usando o logaritmo da variável taxa que é dada por $\text{taxa} = (\text{Número de casos reais} + 2)/\text{Total de habitantes}$. A transformação logarítmica foi devido à distribuição assimétrica significativa da variável taxa. Adicionamos o valor "2" no numerador da variável taxa para não ocorrer problemas quando aplicamos a função logarítmica devido a presença de valores iguais a zero em alguns anos pesquisados.

A Tabela 6.2 mostra os modelos ajustados a cada um dos casos de análise.

Tabela 6.2 – Modelos Ajustados a cada um dos Casos de Análise.

Caso	Modelo	Base de Dados		Predição	
		Localizações	Tempo	Localizações	Tempo
1	Høst et al. (1995): EQM - Modelo 1	52	16 (1991-2006)	13	16 (1991-2006)
	Høst et al. (1995): *di* - Modelo 2	52	16 (1991-2006)	13	16 (1991-2006)
	Gneiting (2002) – Modelo 3	52	16 (1991-2006)	13	16 (1991-2006)
2	Niu et al. (2003): EQM - Modelo 4	65	13 (1991-2003)	13	3 (2004-2006)
	Niu et al. (2003): *di* - Modelo 5	65	13 (1991-2003)	13	3 (2004-2006)
	Gneiting (2002) – Modelo 6	65	13 (1991-2003)	13	3 (2004-2006)
3	Høst et al. (1995): EQM - Modelo 7	52	13 (1991-2003)	13	13 (1991-2003)
	Høst et al. (1995): *di* - Modelo 8	52	13 (1991-2003)	13	13 (1991-2003)
	Niu et al. (2003): EQM - Modelo 9	65	9 (1986-1994)	13	3 (2004-2006)
	Niu et al. (2003): *di* - Modelo 10	65	9 (1986-1994)	13	3 (2004-2006)
	Gneiting (2002) – Modelo 11	65	9 (1986-1994)	13	3 (2004-2006)

6.3 Análise: Caso 1

No Caso 1 de análise o objetivo é predizer a incidência de AIDS nos 13 municípios que foram separados do banco de dados original para testar o modelo nos anos de 1991 a 2006. Para isto utilizamos um banco de dados formado por 52 municípios e para cada município temos a informação do logaritmo da taxa de AIDS nos 16 períodos de tempo.

Inicialmente fizemos uma análise exploratória dos dados com o objetivo de se conhecer melhor os dados e auxiliar na implementação dos modelos. A Figura 6.5 mostra o ln(taxa) de AIDS nos 52 municípios ao longo do tempo e podemos observar que existe uma variação grande do ln(taxa) entre os municípios.

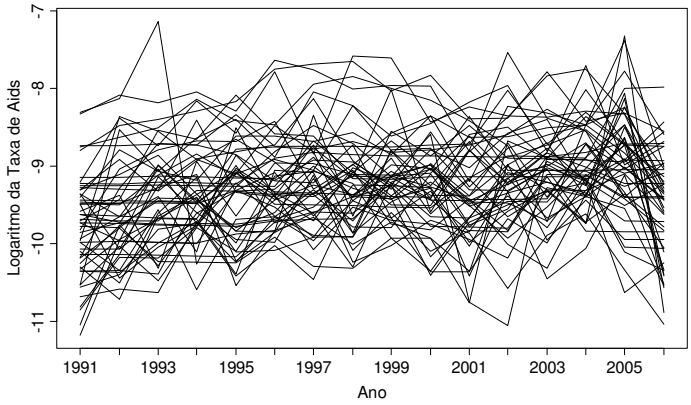


Figura 6.5. Logaritmo da taxa de AIDS nos 52 municípios ao longo do tempo.

Para cada um dos tempos separadamente ajustamos um modelo geoestatístico pelo método de máxima verossimilhança, ainda como uma análise descritiva dos dados. O modelo de variograma ajustado aos dados para todos os anos foi o circular como mostra a Figura 6.6.

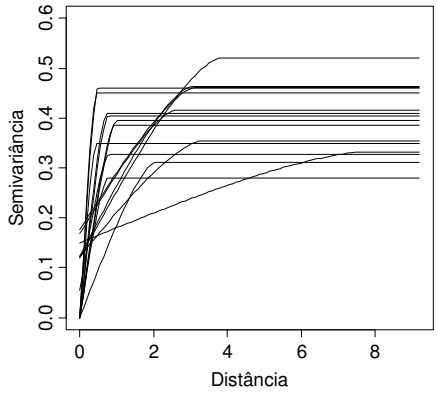


Figura 6.6. Modelo circular ajustado aos dados de ln(taxa) para cada tempo separadamente.

Os parâmetros de média, efeito pepita, variância e de escala estimados pelo modelo circular para cada um dos tempos separadamente são apresentados na Figura 6.7. A linha horizontal representa o valor médio do parâmetro no período de 1991 a 2006. Aparentemente existe uma tendência crescente na componente de média (a) e uma aleatoriedade no comportamento dos outros parâmetros do modelo, efeito pepita (b), variância (c) e escala (d), ao longo do tempo.

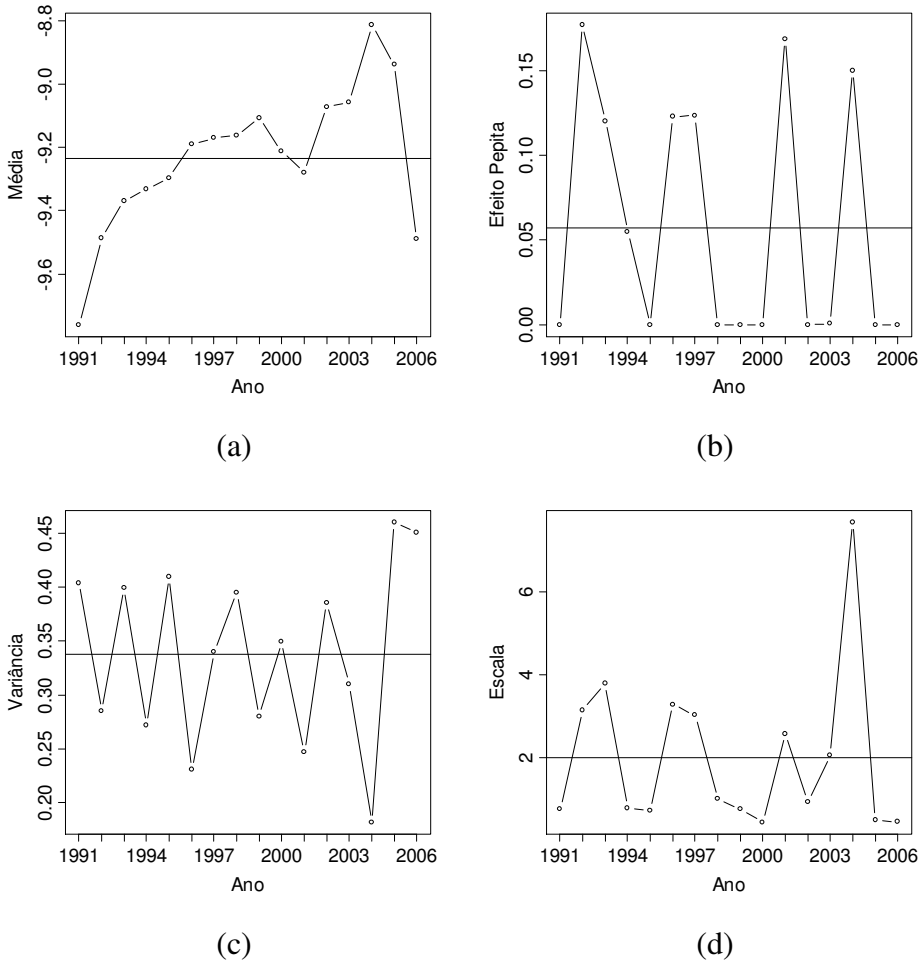


Figura 6.7. Parâmetros estimados para a média (a), efeito pepita (b), variância (c) e escala (d).

O ajuste dos dados pelo modelo proposto por Høst et al. (1995) é apresentado na seção seguinte. Na seção 6.3.2 mostramos o ajuste por funções de covariância da família de Gneiting (2002). Os resultados dos ajustes para o Caso 1 de análise são comparados na seção 6.3.3.

6.3.1 Ajuste pelo Modelo Proposto por Høst et al. (1995)

Nesta seção mostramos o ajuste dos dados pelo modelo proposto por Høst et al. (1995) com os parâmetros estimados pelo método de Kyriakidis e Journel (1999) e com as modificações discutidas no Capítulo 3.

O modelo circular foi ajustado à componente *F* e os parâmetros estimados para a média, o efeito pepita, a variância e a escala são respectivamente: -9,22; 0,03; 0,24 e 3,38, i.e.:

$$\gamma(h)=\begin{cases}0{,}03+0{,}24(\Gamma(h)) & ,h>0\\ 0 & ,h\le 0\end{cases} \tag{6.1}$$

onde $\Gamma(h)=\dfrac{2\left\{\left(\theta\sqrt{1-\theta^2}\right)+sen^{-1}\sqrt{\theta}\right\}}{\pi}$ e $\theta=\min\left(\dfrac{h}{3{,}38},1\right)$.

A Figura 6.8 mostra o modelo de variograma teórico ajustado à componente *F*.

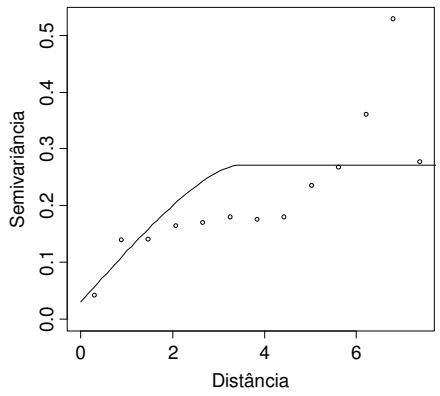


Figura 6.8. Variograma ajustado à componente *F*.

A predição do ln(taxa) de AIDS nos anos de 1991 a 2006 nos 13 municípios de validação foi calculada variando-se a quantidade de vizinhos de acordo com o critério do EQM e da distância *di*. A distância *di* é igual a 3,38 e corresponde ao parâmetro de alcance estimado pelo modelo circular ajustado à componente *F* (ver equação (6.1)). O número de vizinhos baseado no EQM é igual a 10 como mostra a Figura 6.9. O ponto em destaque no gráfico "*" corresponde ao menor valor do EQM. O EQM foi calculado usando as predições nas 13 localidades de validação para os 16 instantes de tempo.

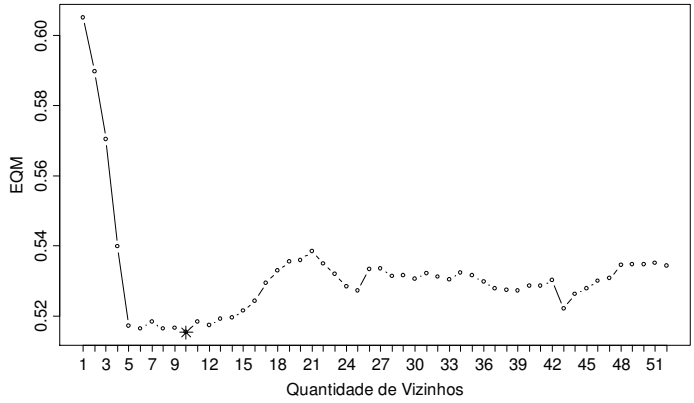


Figura 6.9. EQM de acordo com a quantidade de vizinhos.

A Tabela 6.3 apresenta a quantidade de vizinhos para cada um dos 13 municípios de validação conforme o critério da distância *di*.

Tabela 6.3 – Quantidade de vizinhos de acordo com o município.

Município	Quantidade de Vizinhos
Belo Horizonte	39
Campo Belo	33
Curvelo	39
Diamantina	35
Ipatinga	33
Itaguara	37
Lavras	33
Montes Claros	21
Patrocínio	20
Ponte Nova	32
Pouso Alegre	24
Teófilo Otoni	21
Unaí	8

A Figura 6.10 mostra o EQM de acordo com a distância. O ponto em destaque "*"corresponde a distância *di* igual a 3,38. Este gráfico objetiva avaliar a escolha do valor de *di* e não pode ser construído na prática, pois não temos acesso aos valores reais (ou observados). O EQM foi calculado utilizando as predições nas 13 localidades de validação para os 16 instantes de tempo.

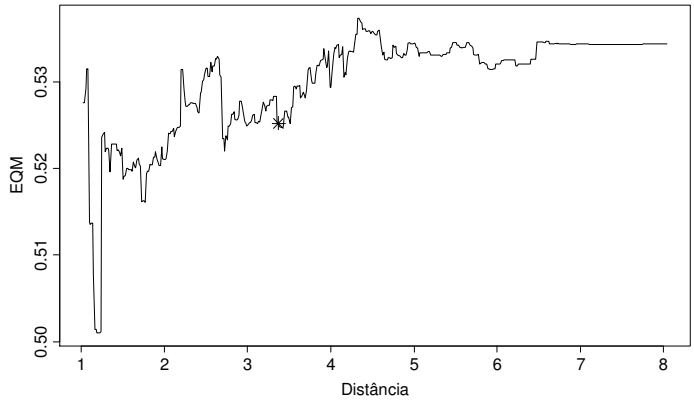


Figura 6.10. EQM de acordo com a distância.

Os modelos ajustados aos dados baseado no EQM e na distância são denotados por respectivamente: Modelo 1 e Modelo 2. O ajuste pelos modelos foi considerado adequado pela validação cruzada. Neste procedimento cada uma das 52 localizações foi retirada da base de dados e o ln(taxa) de AIDS foi estimado por estes modelos nos anos de 1991 a 2006. A distribuição dos resíduos calculados pela diferença entre os valores observados e os valores preditos foi aproximadamente normal, e a média foi próxima de zero.

A seção seguinte apresenta o ajuste pelas funções de covariância da família de Gneiting (2002). A comparação do ajuste por estas funções e pelos Modelos 1 e 2 é mostrada na seção 6.3.3.

6.3.2 Ajuste por Funções de Covariância da Família de Gneiting (2002)

As funções de covariância espaço–temporal separável e não-separável (equações (4.2) e (4.3)) da família de Gneiting ajustadas pelo método de máxima verossimilhança utilizando os dados nas 52 localidades e os 16 instantes de tempo (1991-2006) são dadas respectivamente pelas equações (6.2) e (6.3):

$$C(h,u)=\frac{4{,}2067}{\left(0{,}1228|u|^{1{,}4740}+1\right)^{0{,}1633}}\exp\left(-0{,}0109\|h\|^{0{,}7480}\right)+\frac{0{,}1685}{\left(0{,}1228|u|^{1{,}4740}+1\right)^{0{,}1633}} \tag{6.2}$$

$$C(h,u)=\left(\frac{4{,}2067}{\left(0{,}1228|u|^{1{,}4740}+1\right)^{0{,}0216}}\exp\left(-0{,}0109\left[\frac{\|h\|}{\left(0{,}1228|u|^{1{,}4740}+1\right)^{0{,}0108}}\right]^{0{,}7480}\right)\right)\times\frac{1}{\left(0{,}1228|u|^{1{,}4740}+1\right)^{0{,}1633}}+\frac{0{,}1685}{\left(0{,}1228|u|^{1{,}4740}+1\right)^{0{,}1633}} \tag{6.3}$$

O parâmetro β estimado pela equação (6.3) é igual a 0,0216 indicando que existe uma fraca interação entre os processos espacial e temporal. Como o valor do parâmetro β é próximo de zero vamos considerar que os dois processos são independentes e a predição das observações será calculada pelo ajuste da função dada em (6.2) e denotada por Modelo 3.

A Figura 6.11 mostra a verossimilhança condicional e a perfilhada utilizando o logaritmo da taxa.

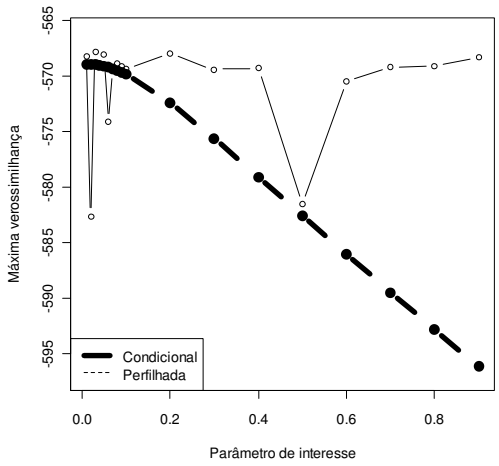


Figura 6.11. Máxima verossimilhança condicional e perfilhada.

A verossimilhança condicional apresenta o maior valor da função em $\beta = 0{,}02$ indicando um modelo de covariância separável. O máximo da verossimilhança perfilhada ocorre quando $\beta = 0{,}03$, porém a diferença entre este valor e aquele quando $\beta = 0$ é pequena. O estudo das verossimilhanças confirma a hipótese de instabilidade do parâmetro β e conseqüentemente a dificuldade na sua estimação.

A seção 6.3.3 compara os modelos 1, 2 e 3 através das predições da taxa de AIDS nas 13 localizações no período de 1991 a 2006.

6.3.3 Comparação dos Modelos Ajustados: Modelos 1, 2 e 3

A Tabela 6.4 mostra o erro quadrático médio (EQM) e a média dos resíduos (RES) pelo ajuste dos Modelos 1, 2 e 3. Os Modelos 1 e 2 se referem a proposta de Høst et al. (1995) que utiliza respectivamente, o critério do erro quadrático médio (EQM) e o critério da distância *di* para a escolha do número de vizinhos para a predição, e o Modelo 3 é concernente à função de covariância separável pertencente à família de Gneiting dada pela

equação (6.2). Estes erros são calculados usando as predições do logaritmo da taxa nas 13 localidades de validação nos 16 instantes de tempo.

Tabela 6.4 – Comparação dos modelos 1, 2 e 3.

Modelo	EQM	RES
Høst et al. (1995): EQM – Modelo 1	0,5153	-0,0134
Høst et al. (1995): *di* - Modelo 2	0,5252	-0,0167
Gneiting (2002): Separável - Modelo 3	0,4875	-0,0141

Os erros associados ao ajuste pelos Modelos 1, 2 e 3 são semelhantes com uma ligeira vantagem para o Modelo 3.

O EQM calculado usando as predições do log da taxa de acordo com o município pode ser visualizado na Figura 6.12. Estes erros são baseados nas predições nas 13 localidades de validação nos anos de 1991 a 2006. Os erros são semelhantes em todos os municípios no que se refere ao modelo ajustado. Os municípios de Belo Horizonte e Itaguara foram os que apresentaram os maiores erros de predição se comparado com as outras localizações.

Para cada ano separadamente foi calculado o EQM usando as predições nos 13 municípios de validação. Os erros de acordo com o ano são apresentados na Figura 6.13. Aparentemente não existe diferença entre os modelos ajustados no que se refere ao EQM, pois o comportamento temporal dos erros é semelhante entre os três modelos. Os maiores erros correspondem aos anos de 1991, 1998 e 2006.

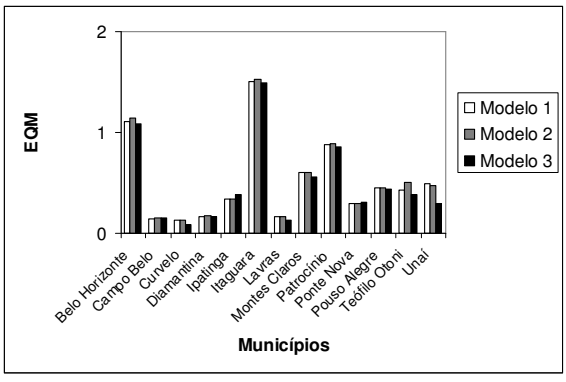


Figura 6.12. Erro quadrático médio por município de acordo com o modelo ajustado.

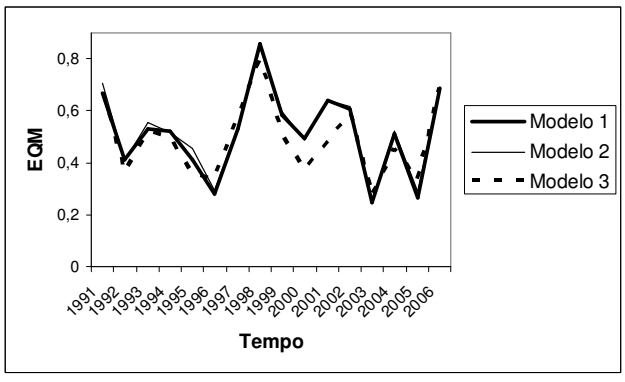


Figura 6.13. Erro quadrático médio por tempo de acordo com o modelo ajustado.

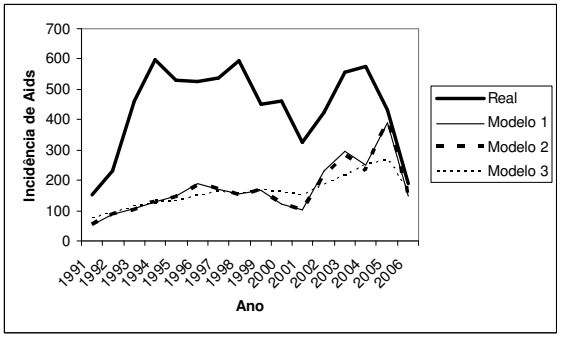
Os quadros de 6.1 a 6.13 apresentam os valores observados e os valores preditos da incidência de AIDS nos 13 municípios de validação para os anos de 1991 a 2006 considerando os Modelos 1, 2 e 3 descritos anteriormente para o ajuste. Os valores preditos

estão na escala original da variável, i.e, aplicou-se a função exponencial as previsões do logaritmo da taxa e estes valores foram multiplicados pela população do ano correspondente e subtraídos de 2 unidades.

Observamos que os menores erros de predição correspondem àqueles municípios onde os vizinhos têm um comportamento semelhante com relação a incidência de AIDS: Campo Belo, Curvelo, Ipatinga e Lavras (ver Figura 6.4). As piores predições correspondem às cidades de Belo Horizonte e Montes Claros, onde em ambas os vizinhos mais próximos apresentam uma quantidade de casos da doença inferior. Aparentemente os modelos ajustados aos dados não conseguem captar variações bruscas do número de casos da doença ao longo do tempo.

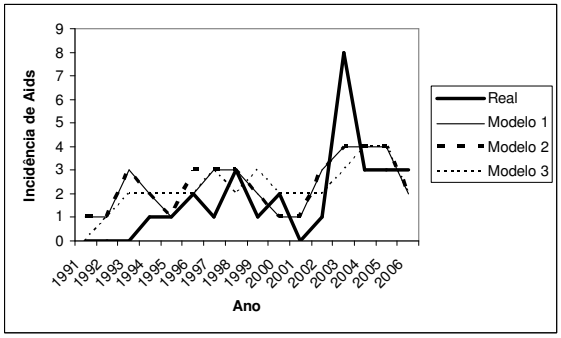
Quadro 6.1. Valores observados e preditos do número de casos de AIDS por ano no município de Belo Horizonte.

Belo Horizonte				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1991	152	52	53	74
1992	231	87	86	92
1993	462	106	103	115
1994	599	128	130	126
1995	528	149	143	134
1996	526	189	185	148
1997	539	170	170	161
1998	595	155	152	155
1999	450	165	166	167
2000	461	122	121	161
2001	326	104	104	150
2002	424	229	217	185
2003	557	295	285	214
2004	576	250	230	249
2005	433	391	395	267
2006	188	149	147	167
Erro médio		269,12	272,50	280,12
$\sqrt{\text{EQM}}$		298,74	302,26	307,70



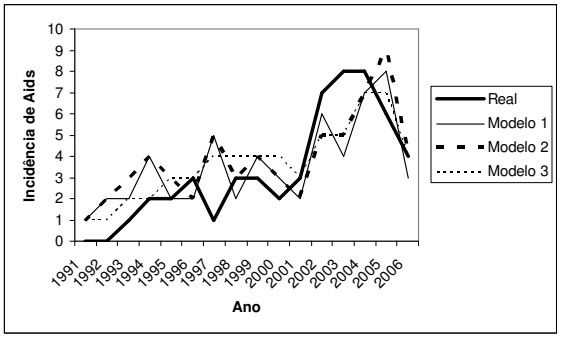
Quadro 6.2. Valores observados e preditos do número de casos de AIDS por ano no município de Campo Belo.

Campo Belo				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1991	0	1	1	0
1992	0	1	1	1
1993	0	3	3	2
1994	1	2	2	2
1995	1	1	1	2
1996	2	2	3	2
1997	1	3	3	3
1998	3	3	3	2
1999	1	2	2	3
2000	2	1	1	2
2001	0	1	1	2
2002	1	3	3	2
2003	8	4	4	3
2004	3	4	4	4
2005	3	4	4	4
2006	3	2	2	2
Erro médio		-0,50	-0,56	-0,44
$\sqrt{\text{EQM}}$		1,62	1,64	1,75



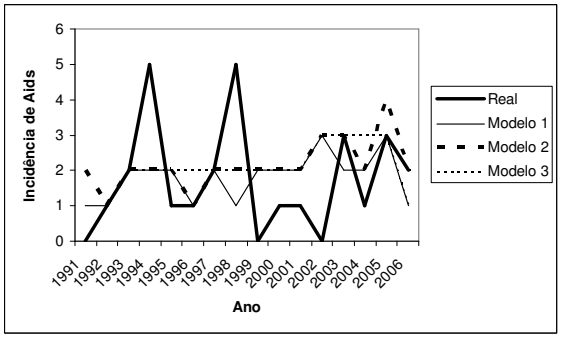
Quadro 6.3. Valores observados e preditos do número de casos de AIDS por ano no município de Curvelo.

Curvelo				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1991	0	1	1	1
1992	0	2	2	1
1993	1	2	3	2
1994	2	4	4	2
1995	2	2	3	3
1996	3	2	2	3
1997	1	5	5	4
1998	3	2	3	4
1999	3	4	4	4
2000	2	3	3	4
2001	3	2	2	3
2002	7	6	5	5
2003	8	4	5	5
2004	8	7	7	7
2005	6	8	9	7
2006	4	3	4	4
Erro médio		-0,25	-0,56	-0,38
$\sqrt{EQM}$		1,84	1,89	1,46



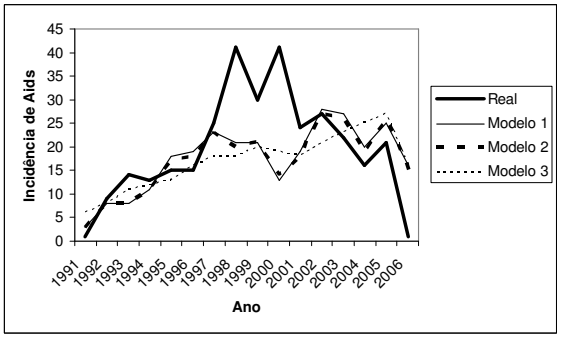
Quadro 6.4. Valores observados e preditos do número de casos de AIDS por ano no município de Diamantina.

Diamantina				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1991	0	1	2	0
1992	1	1	1	1
1993	2	2	2	2
1994	5	2	2	2
1995	1	2	2	2
1996	1	1	1	2
1997	2	2	2	2
1998	5	1	2	2
1999	0	2	2	2
2000	1	2	2	2
2001	1	2	2	2
2002	0	3	3	3
2003	3	2	3	3
2004	1	2	2	3
2005	3	3	4	3
2006	2	1	2	1
Erro médio		-0,06	-0,38	-0,25
$\sqrt{EQM}$		1,68	1,58	1,58



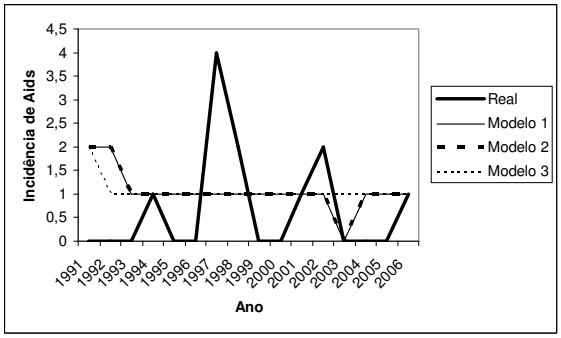
Quadro 6.5. Valores observados e preditos do número de casos de AIDS por ano no município de Ipatinga.

Ipatinga				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1991	1	3	3	6
1992	9	8	8	8
1993	14	8	8	11
1994	13	11	11	12
1995	15	18	17	13
1996	15	19	18	16
1997	25	23	23	18
1998	41	21	20	18
1999	30	21	21	20
2000	41	13	14	19
2001	24	19	18	18
2002	27	28	27	21
2003	22	27	26	23
2004	16	20	19	25
2005	21	25	26	27
2006	1	16	15	16
Erro médio		2,19	2,56	2,75
$\sqrt{EQM}$		10,15	9,98	10,06



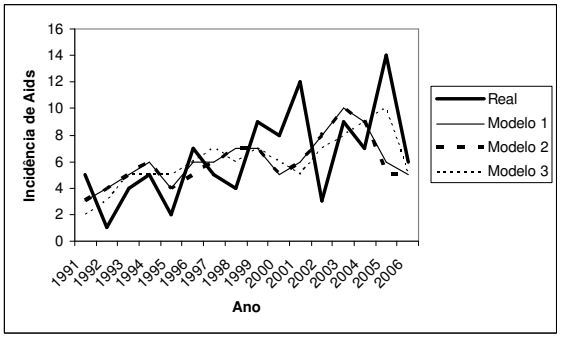
Quadro 6.6. Valores observados e preditos do número de casos de AIDS por ano no município de Itaguara.

Itaguara				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1991	0	0	0	0
1992	0	0	0	0
1993	0	0	0	0
1994	1	0	0	0
1995	0	0	0	0
1996	0	0	0	0
1997	4	0	0	0
1998	2	0	0	0
1999	0	0	0	0
2000	0	0	0	0
2001	1	0	0	0
2002	2	0	0	0
2003	0	0	0	0
2004	0	0	0	0
2005	0	0	0	0
2006	1	0	0	0
Erro médio		1,75	1,75	1,75
$\sqrt{EQM}$		2,09	2,09	2,06



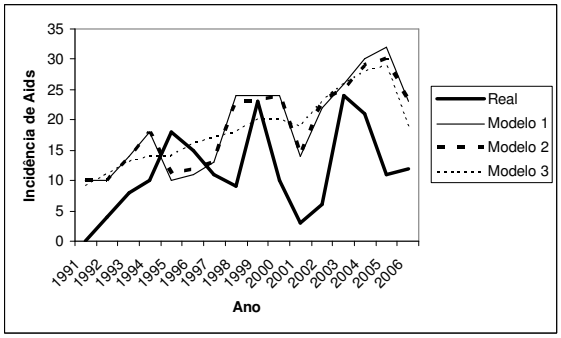
Quadro 6.7. Valores observados e preditos do número de casos de AIDS por ano no município de Lavras.

Lavras				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1991	5	3	3	2
1992	1	4	4	3
1993	4	5	5	5
1994	5	6	6	5
1995	2	4	4	5
1996	7	6	5	6
1997	5	6	6	7
1998	4	7	7	6
1999	9	7	7	7
2000	8	5	5	6
2001	12	6	6	5
2002	3	8	8	7
2003	9	10	10	8
2004	7	9	9	9
2005	14	6	5	10
2006	6	5	5	5
Erro médio		0,25	0,28	0,31
$\sqrt{EQM}$		3,30	3,48	2,82



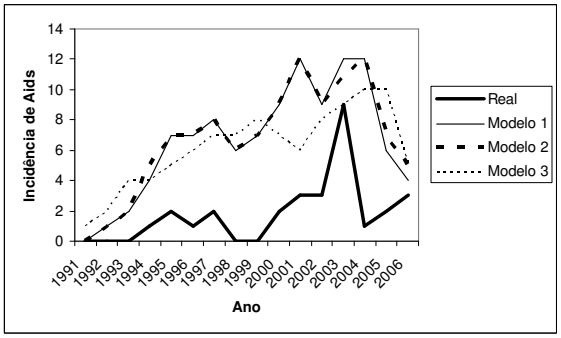
Quadro 6.8. Valores observados e preditos do número de casos de AIDS por ano no município de Montes Claros.

Montes Claros				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1991	0	10	10	9
1992	4	10	10	11
1993	8	14	14	13
1994	10	18	18	14
1995	18	10	11	14
1996	15	11	12	16
1997	11	13	13	17
1998	9	24	23	18
1999	23	24	23	20
2000	10	24	24	20
2001	3	14	14	19
2002	6	22	23	23
2003	24	26	25	26
2004	21	30	29	28
2005	11	32	30	29
2006	12	23	23	19
Erro médio		-7,50	-7,31	-6,94
$\sqrt{EQM}$		10,51	10,15	9,30



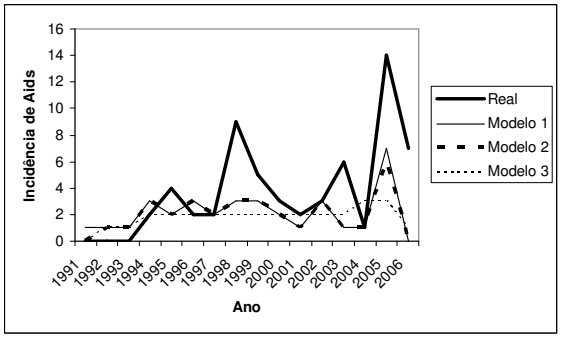
Quadro 6.9. Valores observados e preditos do número de casos de AIDS por ano no município de Patrocínio.

Patrocínio				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1991	0	0	0	1
1992	0	1	1	2
1993	0	2	2	4
1994	1	4	5	4
1995	2	7	7	5
1996	1	7	7	6
1997	2	8	8	7
1998	0	6	6	7
1999	0	7	7	8
2000	2	9	9	7
2001	3	12	12	6
2002	3	9	9	8
2003	9	12	11	9
2004	1	12	12	10
2005	2	6	7	10
2006	3	4	5	5
Erro médio		-4,81	-4,94	-4,38
$\sqrt{EQM}$		5,64	5,72	5,06



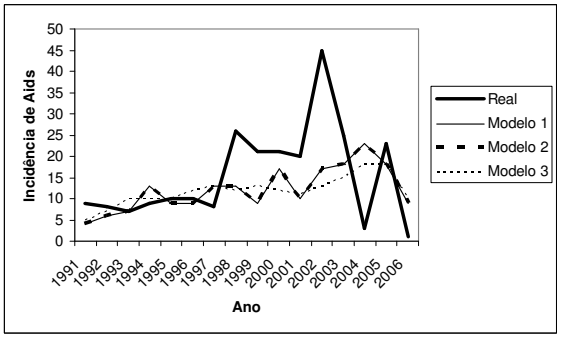
Quadro 6.10. Valores observados e preditos do número de casos de AIDS por ano no município de Ponte Nova.

Ponte Nova				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1991	0	1	0	0
1992	0	1	1	1
1993	0	1	1	1
1994	2	3	3	2
1995	4	2	2	2
1996	2	3	3	2
1997	2	2	2	2
1998	9	3	3	2
1999	5	3	3	2
2000	3	2	2	2
2001	2	1	1	2
2002	3	3	3	2
2003	6	1	1	2
2004	1	1	1	3
2005	14	7	6	3
2006	7	0	0	1
Erro médio		1,62	1,75	1,94
$\sqrt{EQM}$		3,30	3,43	3,90



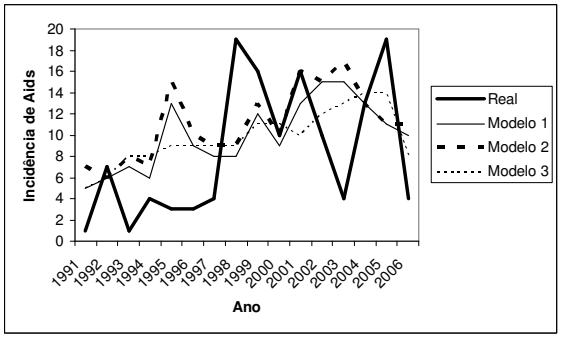
Quadro 6.11. Valores observados e preditos do número de casos de AIDS por ano no município de Pouso Alegre.

Pouso Alegre				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1991	9	4	4	5
1992	8	6	6	7
1993	7	7	7	10
1994	9	13	13	10
1995	10	9	9	10
1996	10	9	9	12
1997	8	13	13	13
1998	26	13	13	12
1999	21	9	9	13
2000	21	17	17	12
2001	20	10	10	11
2002	45	17	17	13
2003	25	18	18	15
2004	3	23	23	18
2005	23	18	18	18
2006	1	9	9	10
Erro médio		3,19	3,19	3,56
$\sqrt{EQM}$		10,67	10,67	10,99



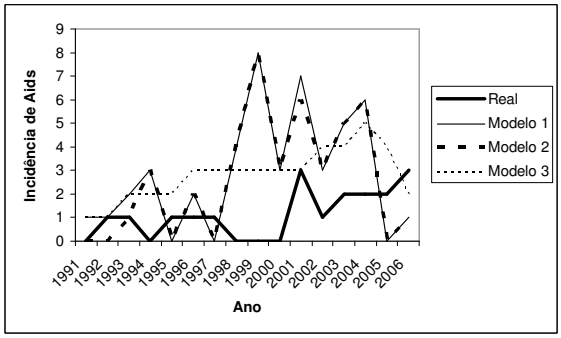
Quadro 6.12. Valores observados e preditos do número de casos de AIDS por ano no município de Teófilo Otoni.

Teófilo Otoni				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1991	1	5	7	5
1992	7	6	6	6
1993	1	7	8	8
1994	4	6	7	8
1995	3	13	15	9
1996	3	9	10	9
1997	4	8	9	9
1998	19	8	9	9
1999	16	12	13	11
2000	10	9	10	11
2001	16	13	16	10
2002	10	15	15	12
2003	4	15	17	13
2004	13	13	13	14
2005	19	11	11	14
2006	4	10	11	8
Erro médio		-1,62	-2,69	-1,38
$\sqrt{EQM}$		6,13	6,75	5,41



Quadro 6.13. Valores observados e preditos do número de casos de AIDS por ano no município de Unaí.

Unaí				
Ano	Real	Modelo 1	Modelo 2	Modelo 3
1991	0	1	0	1
1992	1	1	0	1
1993	1	2	1	2
1994	0	3	3	2
1995	1	0	0	2
1996	1	2	2	3
1997	1	0	0	3
1998	0	4	4	3
1999	0	8	8	3
2000	0	3	3	3
2001	3	7	6	3
2002	1	3	3	4
2003	2	5	5	4
2004	2	6	6	5
2005	2	0	0	4
2006	3	1	1	2
Erro médio		-1,75	-1,50	-1,69
$\sqrt{\text{EQM}}$		3,12	3,04	2,08



A Tabela 6.5 mostra uma análise descritiva dos erros globais associados aos três modelos. Estes erros são calculados pela diferença entre os valores observados e os valores preditos (na escala original da variável) nos 16 instantes de tempo e para os 13 pontos amostrais de validação. Os modelos ajustados parecem razoáveis para predizer a incidência de AIDS nas microrregiões de MG, e aparentemente os erros de predição são semelhantes entre os modelos ajustados.

Na seção seguinte apresenta-se a modelagem do Caso 2 de análise.

Tabela 6.5 – Análise descritiva dos erros.

Método	Média	Média absoluta	Desvio padrão	Mínimo	Máximo
Høst et al. (1995): EQM – Modelo 1	20,125	24,298	80,780	-21	471
Høst et al. (1995): *di* – Modelo 2	20,322	24,543	81,737	-20	469
Gneiting (2002): Separável – Modelo 3	21,154	24,942	83,066	-18	473

6.4 Análise: Caso 2

No Caso 2 de análise o objetivo é predizer a incidência de AIDS nos 13 municípios, excluídos da base de dados original para validação dos modelos, nos anos de 2004, 2005 e 2006. O banco de dados usado neste ajuste tem a informação da variável nos 65 municípios no período de 1991 a 2003. A análise descritiva usando estes dados é muito similar com àquela discutida nas seções 6.1 e 6.2, sendo assim os resultados são suprimidos do texto.

O modelo baseado na proposta de Niu et al. (2003) definido na equação (3.44) e as funções de covariância espaço-temporal separável e não-separável da família de Gneiting (2002) foram ajustados aos dados e os resultados das predições são apresentados nas seções subseqüentes.

6.4.1 Ajuste pelo Modelo Baseado na Proposta de Niu et al. (2003)

Nesta seção mostramos o ajuste dos dados pelo modelo baseado na proposta de Niu et al. (2003) e dado pela equação (3.44). O número de vizinhos do modelo foi determinado por duas formas, sendo a primeira baseada no EQM e a segunda no critério da distância *di*. O EQM é calculado utilizando as predições do logaritmo da taxa de AIDS nos 13 municípios de validação no ano de 2004, e por este critério obtemos uma vizinhança formada pelos 58 municípios mais próximos como mostra a Figura 6.14. O modelo ajustado aos dados baseado no EQM é dado na equação (6.4) onde Z apresenta o logaritmo da taxa:

$$\begin{aligned} Z(s,t) = \; & -3{,}53 + 0{,}49 \times Z(s,t-1) + 0{,}08 \times Z(v_1,t-1) + 0{,}09 \times Z(v_2,t-1) \\ & + 0{,}003 \times Z(v_3,t-1) + 0{,}02 \times Z(v_4,t-1) + 0{,}09 \times Z(v_5,t-1) + \ldots + \\ & - 0{,}04 \times Z(v_{57},t-1) - 0{,}03 \times Z(v_{58},t-1), \quad t = 2004, 2005, 2006 \end{aligned} \tag{6.4}$$

Pela análise da equação (6.4) verificamos que a informação dos vizinhos não parece contribuir significativamente na predição das observações, pois a soma dos pesos associados

aos 58 vizinhos é pequena e igual a 0,12. O peso associado a característica de interesse no próprio local de predição no tempo anterior ao da previsão é igual a 0,49.

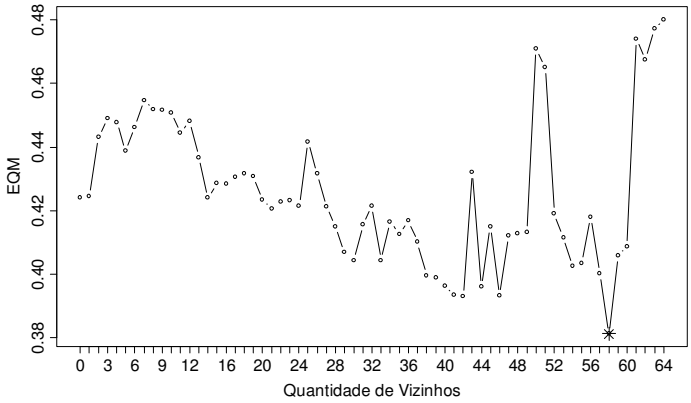


Figura 6.14. EQM de acordo com o número de vizinhos.

A Tabela 6.6 mostra o erro quadrático médio (EQM), a média dos resíduos (RES) e a média da soma de quadrados do erro (MSQE) pelo ajuste do modelo dado na equação (6.4) aos dados. Estes erros são obtidos pelas predições do logaritmo da taxa de AIDS nos anos de 2004, 2005 e 2006 nos 13 municípios separados da base de dados original para a verificação da adequabilidade do modelo.

Tabela 6.6 – Resultados da predição pelo ajuste do modelo dado em (6.4).

Tempo	EQM	RES	MSQE
2004	0,38	-0,37	0,1746
2005	0,31	-0,05	----
2006	0,91	-0,59	----

O primeiro passo (ou etapa) para o cálculo da quantidade de vizinhos baseado no critério da distância *di* é o ajuste de um modelo de variograma teórico aos dados no ano de 2003 (ano anterior ao primeiro de predição). O modelo ajustado aos dados foi o cúbico e os parâmetros estimados para as componentes de média, efeito pepita, variância e alcance são dados por respectivamente: -9,08; 0,10; 0,21 e 3,12, i.e.:

$$\gamma(h)=\begin{cases} 0 & ,h\le 0 \\ 0{,}10+0{,}21\left[7\left(\dfrac{h}{3{,}12}\right)^2-8{,}75\left(\dfrac{h}{3{,}12}\right)^3+3{,}5\left(\dfrac{h}{3{,}12}\right)^5-0{,}75\left(\dfrac{h}{3{,}12}\right)^7\right] & ,0<h\le 3{,}12 \\ 0{,}10+0{,}21 & ,h>3{,}12 \end{cases} \quad (6.5)$$

O variograma cúbico ajustado aos dados no ano de 2003 é mostrado na Figura 6.15.

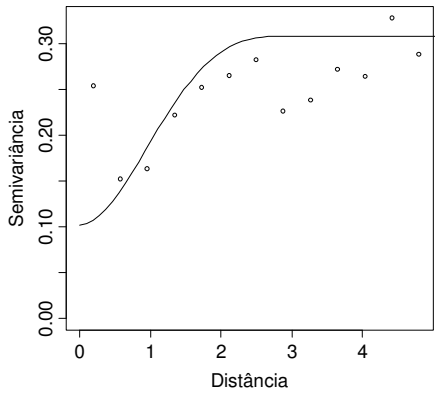


Figura 6.15. Variograma teórico ajustado aos dados no ano de 2003.

O valor da distância *di* é igual a 3,12 e corresponde ao parâmetro de alcance estimado pelo variograma cúbico ajustado aos dados em 2003 (ver equação (6.5)). A Figura 6.16 mostra o EQM de acordo com a distância e podemos observamos que a distância *di* igual a 3,12 (marcado com "*" no gráfico) é uma escolha razoável. O erro é calculado usando as predições nas 13 localizações de validação no ano de 2004 (primeiro tempo de predição).

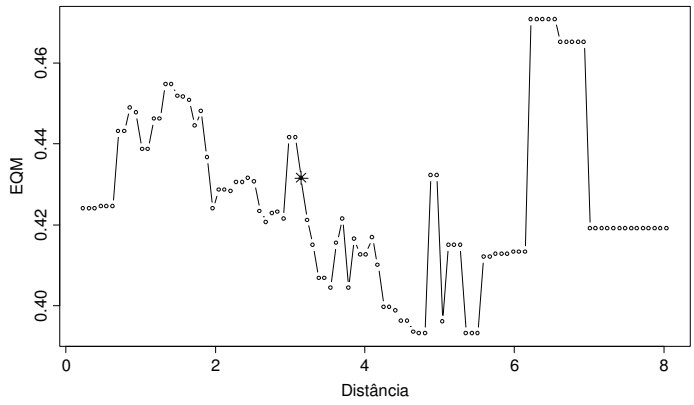


Figura 6.16. EQM de acordo com a distância.

O modelo baseado na distância *di* ajustado aos dados pode ser escrito como:

$$\begin{aligned} Z(s,t) = \; & -2{,}27 + 0{,}57 \times Z(s,t-1) + 0{,}06 \times Z(v_1,t-1) + 0{,}08 \times Z(v_2,t-1) \\ & + 0{,}02 \times Z(v_3,t-1) + 0{,}02 \times Z(v_4,t-1) + 0{,}08 \times Z(v_5,t-1) + \ldots + \\ & + 0{,}05 \times Z(v_{25},t-1) + 0{,}002 \times Z(v_{26},t-1), \quad t = 2004, 2005, 2006 \end{aligned} \tag{6.6}$$

Aparentemente a informação dos vizinhos na predição das observações não é significativa, pois a soma dos pesos associados aos 26 vizinhos é pequena e igual a 0,187. O

logaritmo da taxa de AIDS no local de predição no ano anterior ao da previsão é importante, visto que o peso correspondente a este valor é igual a 0,57.

A descrição dos erros associados ao ajuste do modelo dado na equação (6.6) aos dados é mostrada na Tabela 6.7, conforme o ano de predição (2004, 2005 e 2006).

Tabela 6.7 – Resultados da predição pelo ajuste do modelo dado em (6.6).

Tempo	EQM	RES	MSQE
20	0,43	-0,36	0,1890
21	0,27	-0,05	----
23	0,98	-0,60	----

Os modelos apresentados em (6.4) e (6.6) são denotados por respectivamente: Modelo 4 e Modelo 5. Comparando as Tabelas (6.4) e (6.5) observamos que as diferenças entre os Modelos 4 e 5 são pequenas.

A seção seguinte mostra o ajuste pelas funções de covariância separável e não-separável da família de Gneiting (2002).

6.4.2 Ajuste por Funções de Covariância da Família de Gneiting (2002)

O modelo separável mostrado na equação (4.2) estimado pelo método de máxima verossimilhança usando os dados dos 65 municípios nos anos de 1996 a 2003 é dado por:

$$C(h,u) = \frac{4{,}3733}{\left(0{,}0389|u|^{1{,}9050}+1\right)^{2{,}0167}} \exp\left(-0{,}0109\|h\|^{0{,}6802}\right) + \frac{0{,}1463}{\left(0{,}0389|u|^{1{,}9050}+1\right)^{2{,}0167}} \qquad (6.7)$$

O modelo não-separável ajustado aos dados reduziu-se ao modelo separável da equação (6.7), pois o valor do parâmetro β estimado é igual a zero. O modelo separável é denotado por Modelo 6.

As funções de verossimilhança condicional e perfilhada são apresentadas na Figura 6.17 onde observamos a instabilidade do parâmetro β e a dificuldade na sua estimação.

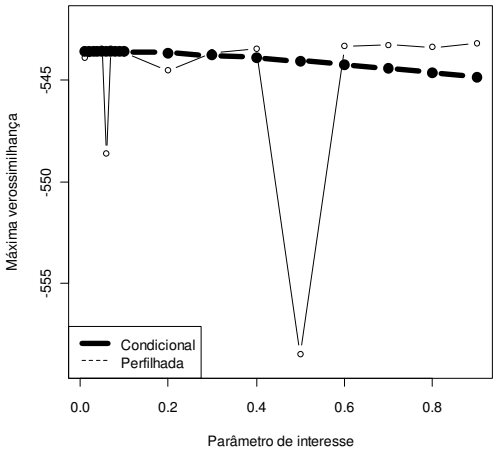


Figura 6.17. Máxima verossimilhança condicional e perfilhada.

A próxima seção compara os ajustes feitos pelos Modelos 4, 5 e 6 aos dados.

6.4.3 Comparação dos Modelos Ajustados: Modelos 4, 5 e 6

A Tabela 6.8 compara os ajustes dos dados pelos Modelos 4, 5 e 6, sendo os dois primeiros modelos baseados na proposta de Niu et al. (2003) e que utilizam respectivamente o critério do EQM e a distância *di* para o cálculo do número de vizinhos e o Modelo 6, se refere ao ajuste pela função de covariância separável da família de Gneiting (2002) dado pela equação (6.7). A tabela apresenta o erro quadrático médio (EQM) e a média dos resíduos (RES) calculados usando as predições do logaritmo da taxa nos 13 municípios de validação nos anos de 2004, 2005 e 2006.

Tabela 6.8 – Comparação dos modelos 4, 5 e 6.

Modelo	EQM	RES
Niu et al. (2003): EQM - Modelo 4	0,53	-0,34
Niu et al. (2003): *di* - Modelo 5	0,56	-0,34
Gneiting (2002): Separável - Modelo 6	0,97	-0,66

Os ajustes pelos Modelos 4 e 5 parecem mais adequados, pois os erros são menores se comparados com o ajuste pelo Modelo 6.

A Figura 6.18 apresenta o erro quadrático médio calculado com as predições do log da taxa nos anos de 2004, 2005 e 2006 de acordo com o município de validação para os Modelos 4, 5 e 6. Os maiores erros correspondem às cidades de Ipatinga e Pouso Alegre para os três

modelos de predição. A Figura 6.19 mostra o EQM por ano para os três modelos ajustados aos dados. O comportamento dos erros nos anos de 2004 e 2005 é semelhante entre os modelos. No ano de 2006 o Modelo 6 apresenta erros maiores se comparado com os Modelos 4 e 5. A pior predição é no ano de 2006 como esperávamos, pois esta predição é baseada em predições de dois anos anteriores.

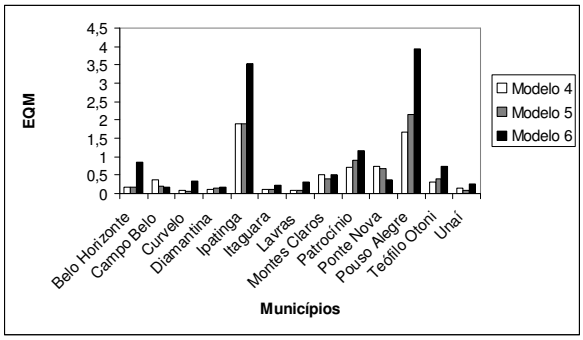


Figura 6.18. Erro quadrático médio por município de acordo com o modelo ajustado.

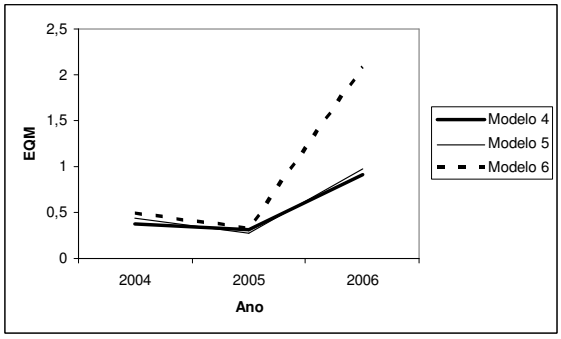


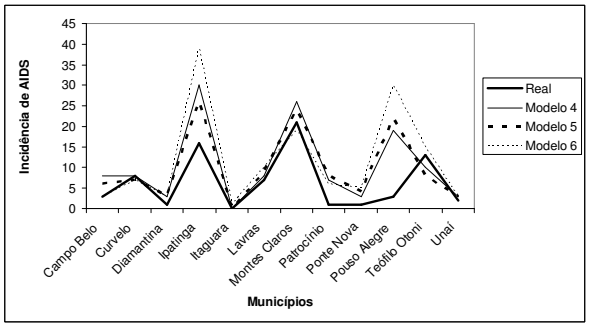
Figura 6.19. Erro quadrático médio por ano de acordo com o modelo ajustado.

Os Quadros de 6.14 a 6.16 apresentam os valores observados e os valores preditos pelos modelos 4, 5 e 6 da incidência de AIDS em cada um dos 13 municípios de validação nos anos de 2004, 2005 e 2006. Estes dados estão na escala original da variável, i.e., aplicou-se a função inversa ao logaritmo da taxa e os valores resultantes foram multiplicados pela população do ano correspondente e subtraídos de 2 unidades. Nos gráficos não incluímos o número de casos de AIDS em Belo Horizonte, pois como os valores são elevados isto interfere na escala dos mesmos. Os valores reais e os valores preditos do número de ocorrências da doença para o município de Belo Horizonte é apresentado na Figura 6.20.

Observamos que as melhores predições no ano de 2004 correspondem às cidades de: Curvelo, Diamantina, Itaguara, Lavras, Montes Claros, Teófilo Otoni e Unaí considerando os três modelos ajustados. No ano de 2005 a predição nestas cidades (exceto em Montes Claros) também foi considerada razoável. As predições no ano de 2006 foram ruins, exceto nas cidades de Itaguara e Diamantina.

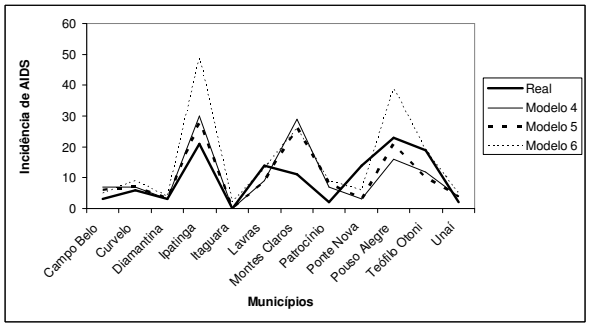
Quadro 6.14. Valores observados e preditos da incidência de AIDS por município no ano de 2004.

Ano - 2004				
Municípios	Real	Modelo 4	Modelo 5	Modelo 6
Belo Horizonte	576	427	436	580
Campo Belo	3	8	6	3
Curvelo	8	8	7	7
Diamantina	1	3	3	3
Ipatinga	16	30	26	39
Itaguara	0	0	0	1
Lavras	7	8	9	10
Montes Claros	21	26	24	19
Patrocínio	1	7	8	6
Ponte Nova	1	3	4	5
Pouso Alegre	3	19	22	30
Teófilo Otoni	13	10	8	15
Unaí	2	3	3	3
Erro médio		7,69	7,38	-5,31
$\sqrt{\text{EQM}}$		41,84	39,39	10,15



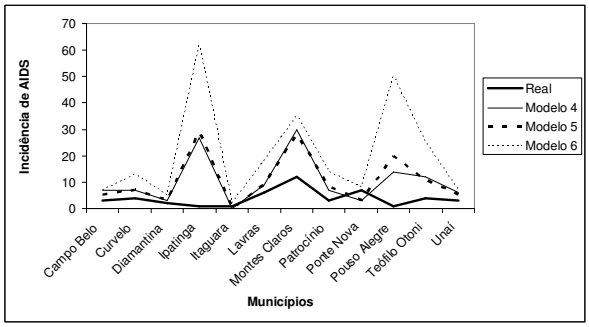
Quadro 6.15. Valores observados e preditos da incidência de AIDS por município no ano de 2005.

Ano - 2005				
Municípios	Real	Modelo 4	Modelo 5	Modelo 6
Belo Horizonte	433	383	386	698
Campo Belo	3	7	6	5
Curvelo	6	7	7	9
Diamantina	3	3	3	4
Ipatinga	21	30	28	49
Itaguara	0	0	0	2
Lavras	14	9	9	13
Montes Claros	11	29	26	26
Patrocínio	2	7	8	9
Ponte Nova	14	3	3	6
Pouso Alegre	23	16	21	39
Teófilo Otoni	19	12	10	19
Unaí	2	4	4	5
Erro médio		3,15	3,08	-25,62
$\sqrt{\text{EQM}}$		15,68	14,58	74,23



Quadro 6.16. Valores observados e preditos da incidência de AIDS por município no ano de 2006.

Ano – 2006				
Municípios	Real	Modelo 4	Modelo 5	Modelo 6
Belo Horizonte	188	358	356	862
Campo Belo	3	7	5	7
Curvelo	4	7	7	13
Diamantina	2	3	3	5
Ipatinga	1	27	29	62
Itaguara	1	0	0	2
Lavras	6	9	9	18
Montes Claros	12	30	28	35
Patrocínio	3	7	8	14
Ponte Nova	7	3	3	8
Pouso Alegre	1	14	20	50
Teófilo Otoni	4	12	11	25
Unaí	3	6	5	7
Erro médio		-19,08	-19,15	-67,15
$\sqrt{\text{EQM}}$		48,21	47,83	188,47



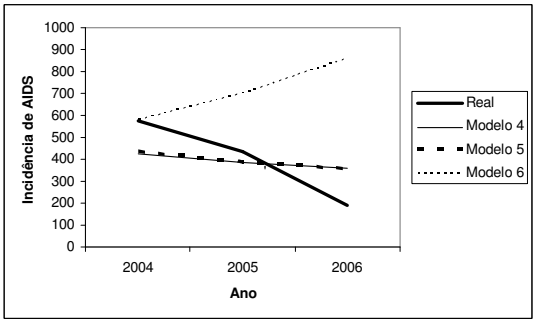


Figura 6.20. Valores observados e preditos da incidência de AIDS em Belo Horizonte nos anos de 2004, 2005 e 2006.

A Tabela 6.9 mostra as medidas de tendência central e de dispersão para descrever os erros globais associados a cada um dos três modelos: 4, 5 e 6. Os erros são calculados pela

diferença entre os valores observados e os valores preditos na escala original da variável nos 13 municípios de validação nos anos de 2004, 2005 e 2006.

Tabela 6.9 – Análise descritiva dos erros.

Método	Média	Média absoluta	Desvio padrão	Mínimo	Máximo
Niu et al. (2003): EQM – Modelo 4	-2,74	14,897	38,344	-170	149
Niu et al. (2003): *di* – Modelo 5	-2,89	14,436	37,117	-168	140
Gneiting (2002): Separável – Modelo 6	-32,69	33,308	113,907	-674	8

Observamos pela Tabela 6.9 que os melhores ajustes correspondem aos Modelos 4 e 5 se comparados com o Modelo 6. Os valores elevados dos erros de predição associados ao ajuste pelo Modelo 6 são devidos, em grande parte, a predição da incidência de AIDS em Belo Horizonte. Se excluirmos o município de Belo Horizonte no cálculo das estatísticas mostradas da Tabela 6.9, observamos que os erros do ajuste pelo Modelo 6 diminui consideravelmente (ver Tabela 6.10), apesar de os ajustes pelos Modelos 4 e 5 ainda serem melhores.

Tabela 6.10 – Análise descritiva dos erros sem o município de Belo Horizonte.

Método	Média	Média absoluta	Desvio padrão	Mínimo	Máximo
Niu et al. (2003): EQM – Modelo 4	-3,78	5,89	7,65	-26	11
Niu et al. (2003): *di* – Modelo 5	-3,67	5,78	7,86	-28	11
Gneiting (2002): Separável – Modelo 6	-9,22	9,89	14,29	-61	8

6.5 Análise: Caso 3

No Caso 3 de análise o objetivo é predizer a incidência de AIDS em municípios e instantes de tempo não observados na amostra. As localizações consideradas desconhecidas para validação de modelos são as mesmas apresentadas anteriormente nos Casos 1 e 2 de análise: Belo Horizonte, Campo Belo, Curvelo, Diamantina, Ipatinga, Itaguara, Lavras, Montes Claros, Patrocínio, Ponte Nova, Pouso Alegre, Teófilo Otoni e Unaí. Os anos considerados na predição do número de casos de AIDS são os de 2004, 2005 e 2006.

O banco de dados usado para testar a adequação dos modelos é formado por 52 municípios e para cada município tem-se a informação da característica de interesse no período de 1991 a 2003. Os resultados da análise descritiva destes dados são similares com aqueles apresentados nas seções 6.1 e 6.2 e por isso não serão reportados no texto.

As metodologias de geoestatística e de séries temporais foram combinadas em duas etapas para predizer estas observações consideradas desconhecidas. Na primeira etapa o modelo de Høst et al. (1995) foi ajustado aos dados com o objetivo de se predizer o logaritmo da taxa de AIDS nos 13 municípios de validação nos anos de 1991 a 2003. Na segunda etapa o banco de dados foi atualizado com as predições calculadas na etapa anterior e ajustou-se o modelo baseado nas idéias de Niu et al. (2003) para predizer o logaritmo da taxa nos anos de 2004, 2005 e 2006 para estas mesmas localidades. Os resultados são mostrados na seção 6.5.1.

As funções de covariância espaço-temporal separável e não-separável (equações (4.2) e (4.3)) da família de Gneiting (2002) também foram ajustadas aos dados para predizer o logaritmo da taxa de AIDS nas 13 cidades de validação nos três últimos anos (2004, 2005 e 2006) e os resultados deste ajuste são apresentados na seção 6.5.2.

6.5.1 Combinação dos Modelos de Geoestatística e de Séries Temporais

Na primeira etapa de predição ajustamos o modelo de Høst et al. (1995) com as modificações descritas na seção 3.5 aos dados, e o modelo de variograma teórico ajustado a componente *F* foi o circular e os parâmetros de média, efeito pepita, variância e de escala estimados pelo método de máxima verossimilhança são respectivamente: -9,25; 0,02; 0,27 e 3,30, i.e.,

$$\gamma(h) = \begin{cases} 0{,}02 + 0{,}27(\Gamma(h)) & , h > 0 \\ 0 & , h \leq 0 \end{cases} \tag{6.8}$$

onde $\Gamma(h) = \dfrac{2\left\{\left(\theta\sqrt{1-\theta^2}\right) + sen^{-1}\sqrt{\theta}\right\}}{\pi}$ e $\theta = \min\left(\dfrac{h}{3{,}30}, 1\right)$.

O variograma circular ajustado aos dados pode ser visualizado pela Figura 6.21.

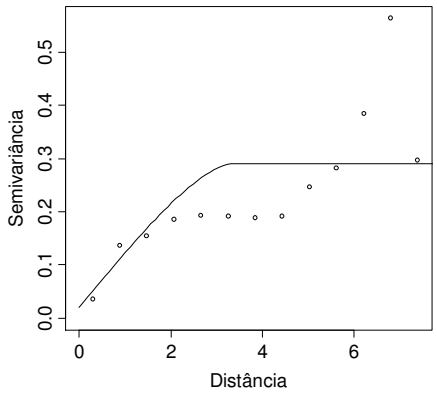


Figura 6.21. Variograma teórico ajustado à componente *F*.

Os critérios do EQM e da distância *di* foram utilizados para o cálculo do número de vizinhos do modelo de Høst et al. (1995). O erro é baseado nas predições do logaritmo da taxa de AIDS nos 13 municípios de validação nos anos de 1991 a 2003, e a distância *di* é igual a 3,30 e corresponde ao parâmetro de alcance estimado pelo modelo circular ajustado à componente *F* (ver equação (6.8)).

Pelo EQM obtemos uma vizinhança formada pelos 5 municípios mais próximos como mostra a Figura 6.22 (ponto "*"), e pela distância *di* o número de vizinhos para cada ponto de predição é dado pela Tabela 6.11. É interessante relembrar que o critério da distância *di* é mais realista que o critério do EQM, visto que este último necessita dos valores reais das observações.

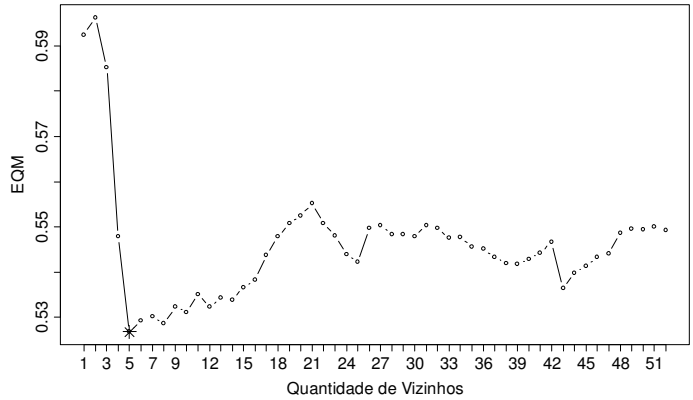


Figura 6.22. EQM de acordo com a quantidade de vizinhos.

Tabela 6.11 – Quantidade de vizinhos de acordo com o município.

Município	Quantidade de Vizinhos
Belo Horizonte	39
Campo Belo	31
Curvelo	37
Diamantina	35
Ipatinga	32
Itaguara	35
Lavras	33
Montes Claros	20
Patrocínio	20
Ponte Nova	31
Pouso Alegre	23
Teófilo Otoni	21
Unaí	8

Para avaliar a escolha da distância *di* construímos o gráfico do EQM de acordo com a distância como mostra a Figura 6.23. O valor da distância *di* igual a 3,30 parece ser razoável (ponto "*"). A construção deste gráfico é inviável em situações reais, pois não temos acesso aos valores verdadeiros da característica de interesse.

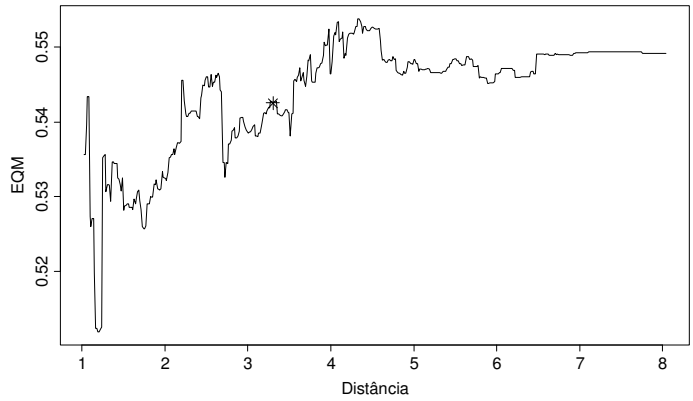


Figura 6.23. EQM de acordo com a distância.

Os modelos ajustados aos dados baseado no EQM e na distância *di* são denotados por respectivamente: Modelo 7 e Modelo 8. Pela validação cruzada verificamos a adequação do ajuste dos dois modelos, visto que os resíduos têm uma distribuição aproximadamente normal e a média dos resíduos é próxima de zero. A Tabela 6.12 resume os erros de predição obtidos pelo ajuste dos Modelos 7 e 8. Os erros são calculados pelas predições feitas nos 13 municípios de validação no período de 1991 a 2003.

Tabela 6.12 – Comparação dos Modelos 7 e 8.

Modelo	EQM	RES
Høst et al. (1995): EQM - Modelo 7	0,5268	0,0218
Høst et al. (1995): *di* - Modelo 8	0,5425	0,0130

A segunda etapa para a predição do logaritmo da taxa de AIDS nos 13 municípios de validação nos anos de 2004, 2005 e 2006 consiste no ajuste do modelo baseado nas idéias de Niu et al. (2003) à base de dados atualizada pelas predições calculadas na primeira etapa. O ajuste do modelo foi feito de duas formas: a primeira baseada no EQM e a segunda na distância *di*. O EQM é calculado com base nas predições no ano de 2004 (primeiro ano de predição) nos 13 municípios de validação, e o valor da distância *di* é igual ao parâmetro de alcance estimado pelo variograma teórico ajustado aos dados no ano de 2003 (ano anterior ao de predição). O modelo de variograma circular foi ajustado aos dados de 2003 e os parâmetros estimados pelo método de máxima verossimilhança para a média, o efeito pepita, a variância e o alcance são dados por respectivamente: -9,06; 0; 0,27 e 2,182, i.e.:

$$\gamma(h) = \begin{cases} 0{,}27(\Gamma(h)) & , h > 0 \\ 0 & , h \leq 0 \end{cases} \tag{6.9}$$

onde $\Gamma(h) = \dfrac{2\left\{\left(\theta\sqrt{1-\theta^2}\right) + sen^{-1}\sqrt{\theta}\right\}}{\pi}$ e $\theta = \min\left(\dfrac{h}{2{,}182}, 1\right)$.

O variograma circular ajustado aos dados neste período é mostrado na Figura 6.24.

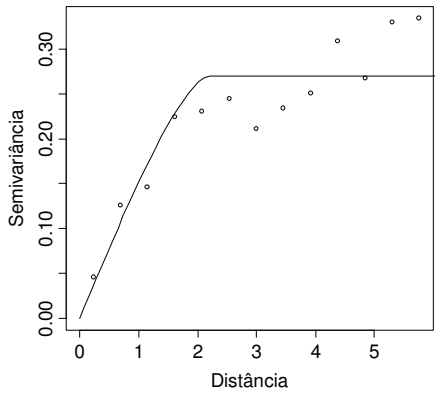


Figura 6.24. Variograma teórico ajustado aos dados no ano de 2003.

O modelo ajustado aos dados pelo critério do EQM é dado pela equação (6.10):

$$
\begin{aligned}
Z(s,t) = \; & -2{,}41 + 0{,}46 \times Z(s,t-1) + 0{,}07 \times Z(v_1,t-1) + 0{,}08 \times Z(v_2,t-1) \\
& + 0{,}001 \times Z(v_3,t-1) + 0{,}07 \times Z(v_4,t-1) + 0{,}07 \times Z(v_5,t-1) + \ldots + \\
& - 0{,}01 \times Z(v_{42},t-1) + 0{,}01 \times Z(v_{43},t-1), \quad t = 2004, 2005, 2006
\end{aligned}
\tag{6.10}
$$

A equação (6.10) diz que a predição em cada um dos 13 municípios de validação nos anos de 2004, 2005 e 2006 utiliza as informações dos 43 vizinhos mais próximos e do próprio local de predição no ano anterior ao da previsão adicionada de uma constante igual a -2,41. Observamos que a contribuição de alguns vizinhos na predição é muito pequena e a soma dos pesos atribuídos a vizinhança é igual a 0,25.

A Tabela 6.13 resume os erros de predição associados ao ajuste do modelo (6.10) aos dados, sendo que o erro quadrático médio (EQM) e a média dos resíduos (RES) foram obtidos pela predição do logaritmo da taxa de AIDS nos 13 municípios de validação para cada um dos anos: 2004, 2005 e 2006.

Tabela 6.13 – Resultados da predição pelo ajuste do modelo dado em (6.10).

Tempo	EQM	RES	MSQE
2004	0,38	-0,33	0,1608
2005	0,40	-0,05	----
2006	1	-0,61	----

O modelo ajustado aos dados usando o critério da distância *di* pode ser escrito como em (6.11):

$$
\begin{aligned}
Z(s,t) = \; & -2{,}19 + 0{,}50 \times Z(s,t-1) + 0{,}07 \times Z(v_1,t-1) + 0{,}08 \times Z(v_2,t-1) \\
& + 0{,}03 \times Z(v_3,t-1) + 0{,}06 \times Z(v_4,t-1) + 0{,}05 \times Z(v_5,t-1) + \ldots + \\
& - 0{,}05 \times Z(v_{15},t-1) - 0{,}04 \times Z(v_{16},t-1), \quad t = 2004, 2005, 2006
\end{aligned}
\tag{6.11}
$$

Neste caso o número de vizinhos utilizados na predição é menor se comparado com o modelo (6.10) e igual a 16. A informação da característica de interesse no local de predição no tempo imediatamente anterior ao da predição é relevante e o peso atribuído a este dado é igual a 0,50. A soma dos pesos associados à vizinhança é igual a 0,24. A Tabela 6.14 mostra os erros de predição pelo ajuste do modelo (6.11) aos dados de acordo com o ano.

Tabela 6.14 – Resultados da predição pelo ajuste do modelo dado em (6.11).

Tempo	EQM	RES	MSQE
2004	0,40	-0,31	0,1667
2005	0,37	-0,01	----
2006	0,97	-0,56	----

Os modelos dados em (6.10) e (6.11) ajustados aos dados são denotados por respectivamente: Modelo 9 e Modelo 10.

Aparentemente não existe diferença entre os Modelos 9 e 10 na predição das observações, exceto no que se refere à quantidade de vizinhos utilizados no ajuste. O modelo 9 utiliza a informação dos 43 vizinhos mais próximos, enquanto que no modelo 10 apenas os 16 locais mais próximos contribuem na predição.

O ajuste dos dados pelos Modelos 9 e 10 serão comparados na seção 6.5.3 com o ajuste dos dados pelas funções de covariância da família de Gneiting. A seção seguinte apresenta o ajuste por estas funções.

6.5.2 Ajuste por Funções de Covariância da Família de Gneiting (2002)

As funções de covariância espaço-temporal separável e não-separável da família de Gneiting apresentadas na seção 4.3.2 ajustadas aos dados pelo método de máxima verossimilhança podem ser escritas como em (6.12) e (6.13) respectivamente:

$$C(h,u)=\frac{4{,}0862}{\left(0{,}0882|u|^{1{,}9533}+1\right)^{0{,}4839}}\exp\left(-0{,}0109\|h\|^{0{,}7517}\right)+\frac{0{,}1472}{\left(0{,}0882|u|^{1{,}9533}+1\right)^{0{,}4839}} \tag{6.12}$$

$$C(h,u)=\left(\frac{4{,}0862}{\left(0{,}0882|u|^{1{,}9533}+1\right)^{0{,}0422}}\exp\left(-0{,}0109\left[\frac{\|h\|}{\left(0{,}0882|u|^{1{,}9533}+1\right)^{0{,}0211}}\right]^{0{,}7517}\right)\right) \times\frac{1}{\left(0{,}0882|u|^{1{,}9533}+1\right)^{0{,}4839}}+\frac{0{,}1472}{\left(0{,}0882|u|^{1{,}9533}+1\right)^{0{,}4839}} \tag{6.13}$$

O valor do parâmetro β na equação (6.13) é igual a 0,0422 e como este valor é muito próximo de zero podemos assumir que os processos espacial e temporal atuam de forma independente.

Pela análise da Figura 6.25 observamos a instabilidade do parâmetro β e a dificuldade em sua estimação.

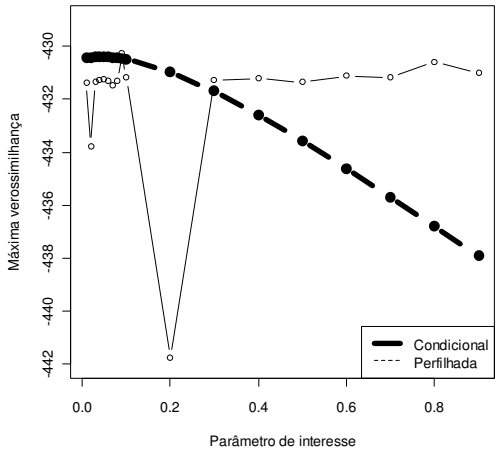


Figura 6.25. Máxima verossimilhança condicional e perfilhada.

O modelo dado pela equação (6.12) ajustado aos dados é denotado por Modelo 11 e será utilizado para predizer as observações nos 13 municípios de validação nos anos de 2004, 2005 e 2006.

A seção seguinte compara os ajustes dos dados pelos Modelos 9, 10 e 11, além de apresentar os valores observados e preditos da incidência de AIDS nos 13 municípios de validação nos três últimos anos.

6.5.3 Comparação dos Modelos Ajustados: Modelos 9, 10 e 11

Nesta seção os Modelos 9, 10 e 11 são comparados utilizando os erros associados às predições do logaritmo da taxa de AIDS nos 13 municípios de validação nos anos de 2004, 2005 e 2006. Os dois primeiros modelos se referem ao ajuste que combina as metodologias de geoestatística e de séries temporais baseado nos critérios do EQM e da distância *di*, respectivamente, para o cálculo do número de vizinhos. O terceiro modelo é relativo ao ajuste dos dados pela função de covariância espaço-temporal separável dada pela equação (6.12). A Tabela 6.15 resume os erros associados ao ajuste para cada um dos três modelos.

Tabela 6.15 – Comparação dos modelos 9, 10 e 11.

Modelo	EQM	RES
Combinado: EQM - Modelo 9	0,59	-0,33
Combinado: *di* - Modelo 10	0,58	-0,29
Gneiting (2002): Separável - Modelo 11	1,04	-0,59

Observamos pela Tabela 6.15 que os Modelos 9 e 10 são preferíveis ao Modelo 11, pois este último fornece erros maiores de predição. Aparentemente não existe diferença entre os Modelos 9 e 10, exceto pelo número de parâmetros estimados pelos modelos como discutido previamente.

A Figura 6.26 apresenta o EQM usando as predições nos anos de 2004, 2005 e 2006 de acordo com o município de validação. Observamos que os maiores erros de predição correspondem às cidades de Ipatinga, Patrocínio e Pouso Alegre. O comportamento dos erros é semelhante entre os modelos ajustados.

A Figura 6.27 mostra o EQM de acordo com o ano, sendo que estes erros são calculados pelas predições do logaritmo da taxa de AIDS nos 13 municípios de validação em cada ano. Os erros associados às predições no ano de 2006 são maiores que para os outros anos considerando os três modelos ajustados. Isto ocorre pelo fato de que as predições no ano de 2006 são baseadas nas predições dos anos anteriores. O Modelo 11 apresenta erros maiores se comparado com os Modelos 9 e 10.

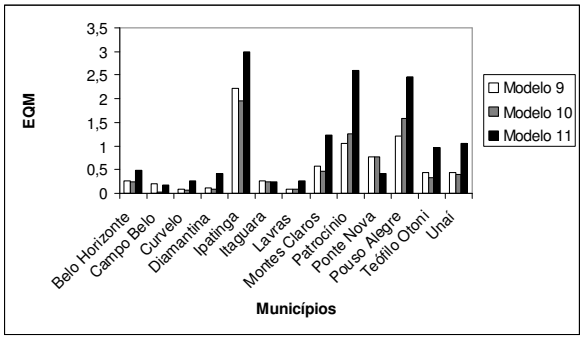


Figura 6.26. EQM de acordo com o município.

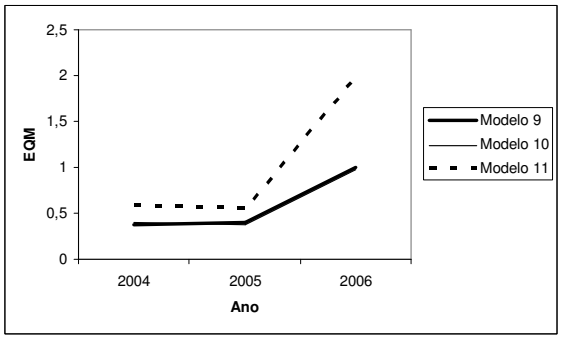


Figura 6.27. EQM de acordo com o ano.

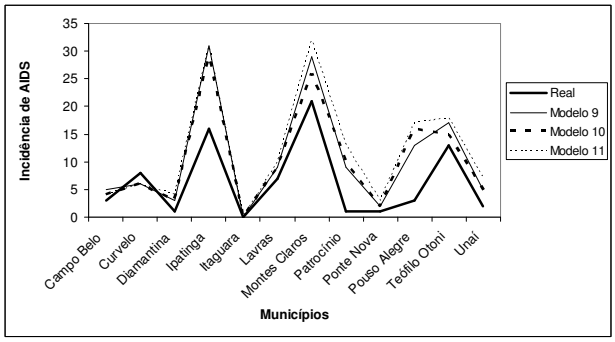
Os Quadros de 6.17 a 6.19 mostram os valores observados e os valores preditos do número de casos de AIDS nos 13 municípios de validação nos anos de 2004, 2005 e 2006. Estes valores estão na escala original da variável, ou seja, aplicou-se a função inversa aos valores preditos e em seguida multiplicou-se pela população do ano correspondente e subtraiu-se 2 unidades.

Os valores reais e preditos do número de ocorrências da doença para o município de Belo Horizonte é apresentado na Figura 6.28, pois os valores elevados da quantidade destes casos prejudicam a visualização do comportamento da incidência de AIDS nos outros municípios.

No ano de 2004 as melhores predições considerando os três modelos correspondem aos municípios de: Campo Belo, Curvelo, Diamantina, Itaguara, Lavras, Ponte Nova, Teófilo Otoni e Unaí. Para o ano de 2005 as predições feitas nestas mesmas cidades, exceto em Ponte Nova e Unaí também foram consideradas boas. As predições no ano de 2006 foram razoáveis apenas nas cidades de Diamantina e Itaguara, onde o número de casos da doença é pequeno e/ou nulo.

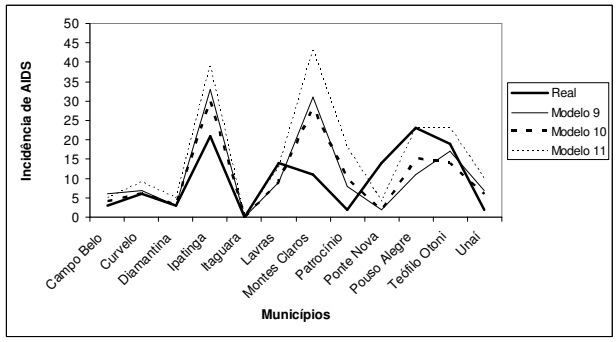
Quadro 6.17. Valores observados e preditos da incidência de AIDS por município no ano de 2004.

Ano - 2004				
Municípios	Real	Modelo 9	Modelo 10	Modelo 11
Belo Horizonte	576	315	306	237
Campo Belo	3	5	4	4
Curvelo	8	6	6	6
Diamantina	1	3	3	4
Ipatinga	16	31	29	31
Itaguara	0	0	0	0
Lavras	7	9	9	10
Montes Claros	21	29	26	32
Patrocínio	1	9	10	13
Ponte Nova	1	2	2	3
Pouso Alegre	3	13	16	17
Teófilo Otoni	13	17	15	18
Unaí	2	5	5	7
Erro médio		16	17	20,77
$\sqrt{EQM}$		72,65	75,13	94,33



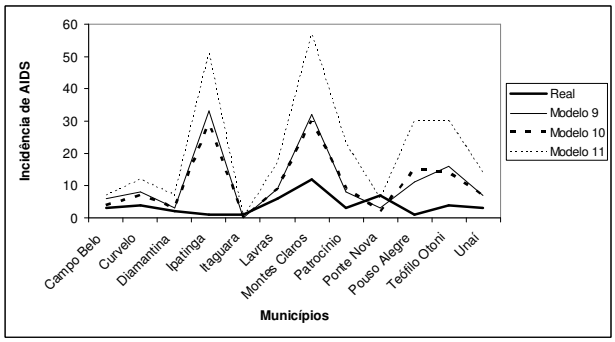
Quadro 6.18. Valores observados e preditos da incidência de AIDS por município no ano de 2005.

Ano – 2005				
Municípios	Real	Modelo 9	Modelo 10	Modelo 11
Belo Horizonte	433	331	306	301
Campo Belo	3	6	4	5
Curvelo	6	7	6	9
Diamantina	3	3	3	5
Ipatinga	21	33	30	39
Itaguara	0	0	0	0
Lavras	14	9	9	13
Montes Claros	11	31	28	43
Patrocínio	2	8	10	18
Ponte Nova	14	2	2	4
Pouso Alegre	23	11	15	23
Teófilo Otoni	19	17	14	23
Unaí	2	7	6	10
Erro médio		6,62	9,08	4,46
$\sqrt{EQM}$		29,53	35,99	38,46



Quadro 6.19. Valores observados e preditos da incidência de AIDS por município no ano de 2006.

Ano – 2006				
Municípios	Real	Modelo 9	Modelo 10	Modelo 11
Belo Horizonte	188	336	301	391
Campo Belo	3	6	4	7
Curvelo	4	8	7	12
Diamantina	2	3	3	7
Ipatinga	1	33	29	51
Itaguara	1	0	0	0
Lavras	6	9	9	17
Montes Claros	12	32	30	57
Patrocínio	3	8	9	23
Ponte Nova	7	3	2	6
Pouso Alegre	1	11	15	30
Teófilo Otoni	4	16	14	30
Unaí	3	7	7	14
Erro médio		-18,23	-15	-31,54
$\sqrt{\text{EQM}}$		42,67	33,13	60,76



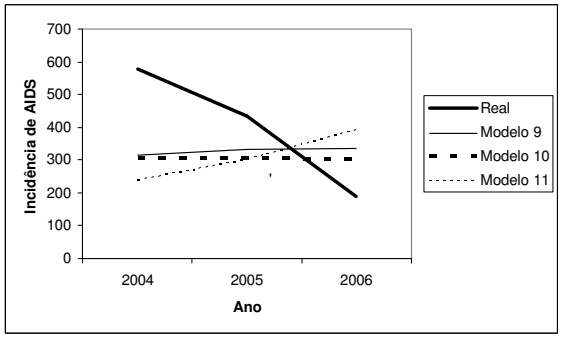


Figura 6.28. Valores observados e preditos da incidência de AIDS em Belo Horizonte nos anos de 2004, 2005 e 2006.

A Tabela 6.16 mostra as medidas de tendência central e de dispersão para descrever os erros globais associados a cada um dos três Modelos: 9, 10 e 11. Os erros são calculados pela

diferença entre os valores observados e os valores preditos na escala original da variável nos 13 municípios de validação nos anos de 2004, 2005 e 2006.

Tabela 6.16 – Análise descritiva dos erros.

Método	Média	Média absoluta	Desvio padrão	Mínimo	Máximo
Combinação: EQM – Modelo 9	1,462	19,103	52,197	-148	261
Combinação: *di* – Modelo 10	3,692	18,615	52,302	-113	270
Gneiting (2002): Separável – Modelo 11	-2,103	27,026	69,346	-203	339

A média dos erros associados ao ajuste é semelhante entre os três modelos considerados. O Modelo 11 apresenta uma maior variação dos erros se comparado com os Modelos 9 e 10.

Aparentemente não existe diferença entre o modelo de Høst et al. (1995) e a função de covariância separável na interpolação espacial do número de casos de AIDS em tempos amostrados. Porém vale ressaltar a instabilidade do parâmetro β e a dificuldade em sua estimação no modelo de covariância não-separável.

As predições da incidência de AIDS em tempos futuros e em localizações e tempos não observados na amostra utilizando respectivamente os modelos de Niu et al. (2003) e a estratégia de combinar os modelos de Høst et al. (1995) e de Niu et al. (2003) forneceram melhores resultados que o ajuste pela função de covariância separável da classe de Gneiting (2002).

No Capítulo seguinte apresentamos as considerações finais sobre os ajustes dos modelos aos bancos de dados reais da taxa de criminalidade, da armazenagem de água em um solo cultivado com citros e da incidência de AIDS em MG, apontando as principais conclusões referentes às vantagens e desvantagens e a viabilidade de implementação destes modelos formadas a partir das análises discutidas nos Capítulos 4, 5 e 6.

Capítulo 7

Considerações Finais

Neste trabalho foram feitas três análises distintas de dados denominadas Casos 1, 2 e 3 de acordo com o objetivo do estudo e com as informações (ou dados) disponíveis. Nessas análises utilizamos dados provenientes de áreas distintas: o primeiro banco de dados trata da taxa de criminalidade nas regiões administrativas do estado de Minas Gerais, o segundo é relacionado a fenômenos ambientais e estuda a armazenagem de água em um solo cultivado com citros, e o terceiro é relativo à incidência de AIDS nas microrregiões de Minas Gerais.

Para a análise dos Casos 1, 2 e 3 foi implementado computacionalmente no *software* R o modelo geoestatístico proposto por Høst et al. (1995) com as componentes estimadas pelo método mostrado em Kyriakidis e Journel (1999) e um modelo de séries temporais baseado nas idéias de Niu et al. (2003).

No Caso 1 de análise o objetivo foi fazer a interpolação espacial da característica de interesse em tempos observados na amostra e dois modelos foram ajustados aos dados com esta finalidade. O primeiro modelo é o de Høst et al. (1995) com as componentes estimadas pelo método proposto por Kyriakidis e Journel (1999) e com algumas modificações referentes a quantidade de vizinhos utilizada na predição descritas no Capítulo 3. O segundo é relativo as funções de covariância espaço-temporal separável e não-separável da classe de Gneiting (2002).

Observamos pelos estudos de caso que não foi necessário utilizar todos os vizinhos na interpolação espacial da característica de interesse quando ajustamos o modelo de Høst et al. (1995) aos dados, ou seja, os erros de predição são menores quando usamos apenas a informação dos vizinhos mais próximos do ponto de predição. Concluímos que a alteração na modelagem relativa ao número de vizinhos usados na predição das observações parece ser adequada.

Todos os processos analisados forneceram valores do parâmetro β referente a função de covariância não-separável da classe de Gneiting (2002) iguais ou próximos de zero, indicando que existe uma fraca interação entre as dimensões espacial e temporal. Desta forma o modelo não-separável se reduz ao modelo separável e consideramos que as estruturas espaço e tempo são independentes. Segundo Huang et al. (2007, p. 4594) “a estrutura separável fornece bons resultados de predição mesmo para dados gerados de estruturas não-

separáveis. Isto se deve simplesmente a flexibilidade intrínseca dos modelos separáveis propostos". Vale ressaltar que devido a dificuldades nos algoritmos numéricos na estimação do parâmetro β foram adotadas duas estratégias para contornar este problema. O estudo das verossimilhanças (condicional e perfilhada) também indicou em ambos os estudos de casos valores deste parâmetro próximo de zero.

Pelos resultados das análises de dados do Caso 1 notamos que os valores preditos são mais próximos dos valores reais quando a variável que está sendo medida apresenta um comportamento semelhante nos vizinhos e no ponto de predição, i.e., quanto maior a similaridade da característica avaliada entre o ponto em que se deseja fazer a predição e os vizinhos, menores são os erros de predição. Observamos também que o ajuste desses modelos é mais favorável a dados relacionados a fenômenos ambientais em relação a dados oriundos de outras ciências como biologia e sociologia, pois estes últimos apresentam grande variação e ainda existem diversos fatores que influenciam a característica medida que não podem ser controlados ou nem mesmo são conhecidos. Isto justifica em parte a grande quantidade de artigos da área de processos espaço-temporais que trabalham com dados relacionados a fenômenos ambientais.

Os erros de predição provenientes do ajuste dos dados pelo modelo de Høst et al. (1995) são menores, na maioria dos casos, que aqueles obtidos pelo ajuste da função de covariância separável da família de Gneiting (2002) indicando que o modelo de Høst et al. (1995) é uma alternativa prática as funções de covariância de Gneiting (2002).

No Caso 2 o propósito da análise foi fazer a previsão temporal em pontos observados na amostra, e para isso ajustou-se o modelo baseado nas idéias de Niu et al. (2003) e as funções de covariância da família de Gneiting (2002).

Na modelagem baseada na proposta de Niu et al. (2003) observamos a importância (ou influência) dos vizinhos no cálculo das predições, pois em todos os estudos de casos tratados neste trabalho o modelo considerou as informações dos vizinhos nas predições.

Os erros obtidos pelo ajuste do modelo de Niu et al. (2003) foram menores ou similares aqueles resultantes pelo ajuste da função de covariância de Gneiting (2002), sendo assim o modelo baseado na idéias de Niu et al. (2003) e dado pela equação (3.44) é preferível as funções de covariância da classe de Gneiting (2002) dadas pelas equações (4.2) e (4.3).

No Caso 2 de análise também observamos que os erros de predição estão associados com o comportamento da característica de interesse na vizinhança do ponto que se deseja fazer a predição, além disso, notamos que processos mais estáveis, i.e., processos os quais a

variação entre as observações coletadas ao longo do tempo é pequena resultam em erros menores de predição.

No desenvolvimento deste trabalho testamos uma proposta alternativa para a previsão temporal das observações em pontos amostrados onde o modelo de Niu et al. (2003) era sempre atualizado com as novas predições, ou seja, para cada tempo de predição tínhamos um modelo diferente e as componentes para cada modelo eram reestimadas utilizando a informação das predições anteriores. Esta modelagem forneceu resultados muito semelhantes a aqueles apresentados no texto e, portanto preferimos omitir os resultados dessa análise nesta dissertação.

No Caso 3 de análise combinamos a metodologia de geoestatística e de séries temporais ajustando os modelos propostos por Høst et al. (1995) e por de Niu et al. (2003) em duas etapas com o objetivo de se predizer as observações em localizações e tempos não amostrados. As funções de covariância da classe de Gneiting (2002) também foram ajustadas com esse propósito. As conclusões são similares a aquelas discutidas previamente e observamos que para os dados utilizados nesta dissertação as predições usando esta estratégia de combinar esses dois modelos forneceram melhores resultados, na maioria dos casos, se comparados com o ajuste pelas funções de covariância de Gneiting (2002).

Nos Casos 1, 2 e 3 de análise aplicando os modelos de Høst et al. (1995), o de Niu et al. (2003) e a combinação de ambos, o critério da distância *di* forneceu resultados bons de predição se comparado com o método pelo erro quadrático médio (EQM) que foi usado para efeito de comparação e podemos concluir que a distância *di* é um critério razoável na quantificação de vizinhos para a predição.

A representação dos processos espaço-temporais por funções de covariância, embora seja promissor e recentemente dispormos de diversos trabalhos na literatura que se preocupam com a proposição de funções de covariância válidas, apresenta dificuldades e restrições na implementação computacional, e conseqüentemente prejudica ou inviabiliza a análise de dados reais indexados pelo espaço e pelo tempo. Sendo assim, os modelos de Høst et al. (1995) e de Niu et al. (2003) são alternativas práticas e acessíveis na predição de observações espaço-temporais, especialmente nos casos em que a característica de interesse no ponto de predição e nos vizinhos é similar e/ou em situações onde a série de observações é estável ao longo do tempo.

Finalizamos nossas conclusões ressaltando que um dos fatores penosos no desenvolvimento deste trabalho foi a imensa dificuldade em se conseguir banco de dados, pois a maioria dos pesquisadores e/ou das instituições não disponibilizam os mesmos. A idéia

inicial do estudo era aplicar esta metodologia a uma variedade de dados, porém diante dessas limitações tivemos que adequar os nossos objetivos a realidade.

7.1 Sugestões e Trabalhos Futuros

Pretendemos disponibilizar os modelos implementados computacionalmente nesta dissertação em um pacote no *software* R para que as pessoas interessadas em fazer as análises espaço-temporais possam usufruir do produto (ou programa) gerado por este estudo.

A seguir enumeramos algumas sugestões para trabalhos futuros:

- Expandir o modelo baseado nas idéias de Niu et al. (2003) introduzindo um modelo de média móvel (MA) ou um autoregressivo média móvel (ARMA) a essa proposta.
- Estimar o desvio-padrão dos erros de predição e construir intervalos de confiança para os modelos considerados no estudo.
- Utilizar outros critérios além da distância *di* para a obtenção do número de vizinhos na predição considerando os modelos de Høst et al. (1995) e de Niu et al. (2003).
- Utilizar outras medidas de distância como a *great circle distances* que leva em consideração a curvatura da Terra, e deve ser útil em situações onde os dados são separados entre si por grandes distâncias.

BIBLIOGRAFIA

1. ALMEIDA, C. F. P.; RIBEIRO JÚNIOR, P. J. *Estimativa da distribuição espacial de retenção de água em um solo utilizando krigagem indicatriz*. Curitiba, 1996. 36p. (Relatório Técnico do Laboratório de Estatística).

2. ARMSTRONG, M. *Basic linear geostatistics.* Berlin: Springer, 1998. 153p.

3. BAILEY, T. C.; GATRELL, A. C. *Interactive spatial data analysis*. London: Longman, 1995.

4. BANERJEE, S.; CARLIN, B. P.; GELFAND, A. E. *Hierarchical modeling and analysis of spatial data*. New York: Chapman and Hall, 2004.

5. BEATO FILHO, C. C.; ASSUNÇÃO, R.; SANTOS, M. A. C.; SANTO, C. L. E. E.; SAPORI, L. F.; BATITUCCI, E.; MORAIS, P. C. C.; SILVA, S. L. F. D. Criminalidade violenta em Minas Gerais – 1986 a 1997. In: ASSOCIAÇÃO NACIONAL DE PÓS-GRADUAÇÃO E PESQUISA EM CIÊNCIAS SOCIAIS, XXII, 1998, Caxambu, p. 1-28.

6. BOCHNER, S. *Harmonic analysis and theory of probability*. Berkeley and Los Angeles: University of California Press, 1955.

7. BOX, G. E. P.; COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society*, Series B (Methodological), v. 26, n. 2, p. 211-252, 1964.

8. BROWN, P. E.; DIGGLE, P. J.; LORD, M. E.; YOUNG, P. C. Space time calibration of radar rainfall data. *Applied Statistics*, vol. 50, part 2, p. 221-241, 2001.

9. CHILÉS, J. P.; DELFINER, P. *Geostatistics modeling spatial uncertainty*. New York: John Wiley & Sons, 1999. 695p.

10. CRESSIE, N. *Statistics for spatial data*. Revised Edition. New York: John Wiley & Sons, 1993. 900p.

11. CRESSIE, N.; CHAN, N. H. Spatial modeling of regional variables. *Journal of the American Statistical Association*, vol. 84, n. 406, p. 393-401, 1989.

12. CRESSIE, N.; HAWKINS, D. M. Robust estimation of the variogram, I. *Journal of the International Association for Mathematical Geology*, 12, p. 115-125, 1980.

13. CRESSIE, N.; HUANG, H.-C. Classes of nonseparable, spatio–temporal stationary covariance functions. *Journal of the American Statistical Association*. V. 94, n. 448, p. 1330-1340, dez, 1999.

14. CURRIERO, F. C.; LELE, S. A composite likelihood approach to semivariogram estimation. *Journal of Agricultural, Biological, and Environmental Statistics,* vol. 4, n.1, p. 9-28, 1999.

15. DE IACO, S.; MYERS, D. E.; POSA, D. Nonseparable space-time covariance models: some parametric families. *Mathematical Geology,* vol. 34, n. 1, p. 23-42, 2002.

16. _____. The linear coregionalization model and the product-sum space-time variogram. *Mathematical Geology*, vol. 35, n. 1, p. 25-38, 2003.

17. DE LUNA, X.; GENTON, M. G. Predictive spatio-temporal models for spatially sparse environmental data. Statistica Sinica, vol. 15, p. 547-568, 2005.

18. DIGGLE, P. J. *Statistical analysis of spatial point patterns*. London: Arnold, 2003.

19. DIGGLE, P. J.; RIBEIRO JÚNIOR, P. J. *Model-based geostatistics*. New York: Springer, 2007. 228p.

20. FRANÇA, J. L. *Manual para normalização de publicações técnico-científicas*. 7 ed. Belo Horizonte: Ed. UFMG, 2004. 242p. (Aprender; 15).

21. GENTON, M. G. Highly robust variogram estimation. *Mathematical Geology*, vol. 30, n. 2, p. 213-221, 1998.

22. geoR: a package for geostatistical analysis, RIBEIRO JÚNIOR, P. J.; DIGGLE, P.J. , *R-News*, v. 1, n. 2, p. 14-18, 2001, http://CRAN.R-project.org/doc/Rnews/.

23. GNEITING, T. Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, v. 97, n. 458, p. 590-600, jun, 2002.

24. GNEITING, T.; SCHLATHER, Martin. Space–time covariance models. *Encyclopedia of Environmetrics*, v. 4, p. 2041-2045, 2002.

25. GOOVAERTS, P. *Geostatistics for natural resources evaluation*. New York: Oxford, 1997. 483p. (Applied Geostatistics Series).

26. HANDCOCK, M. S.; WALLIS, J. R. An Approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association*, vol. 89, n. 426, p. 368-378, 1994.

27. HASLETT, J. On the sample variogram and the sample autocovariance for non-stationary time series. *The Statistician*, vol. 46, p. 475-485, 1997.

28. HASLETT, J.; RAFTERY, A. E. Space-time modelling with long-memory dependence: assessing ireland´s wind power resource. (with discussion). *Applied Statistics*, vol. 38, p. 1-50, 1989.

29. HØST, G.; OMRE, H.; SWITZER, P. Spatial interpolation errors for monitoring data. *Journal of the American Statistical Association*, vol. 90, n. 431, p. 853-861, 1995.

30. HUANG, H.-C.; MARTINEZ, F.; MATEU, J.; MONTES, F. Model comparison and selection for stationary space–time models. *Computational Statistics & Data Analysis*. N. 51, p. 4577-4596, ago, 2007.

31. HUERTA, G.; SANSÓ, B.; STROUD, J. R. A spatiotemporal model for mexico city ozone levels. *Applied Statistics*, vol. 53, part 2, p. 231-248, 2004.

32. JOURNEL, A. G. New distance measures: The route toward truly non-gaussian geostatistics. *Mathematical Geology*, vol.2, n. 04, p. 459-475, 1988.

33. JOURNEL, A. G.; HUIJBREGTS, CH. J. *Mining geostatistics*. New York: Academic Press, 1978. 600p.

34. KYRIAKIDIS, P. C.; JOURNEL, A. G. Geostatistical space-time models: a review. *Mathematical Geology*, vol. 31, n. 06, p. 651-684, 1999.

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

36. MA, C. Spatio–temporal covariance functions generated by mixtures. *Mathematical Geology*, v. 34, n. 8, p.965-975, nov, 2002.

37. _____. Families of spatio–temporal stationary covariance models. *Journal of Statistical Planning and Inference*, n. 116, p. 489-501, mai, 2003.

38. _____. Linear combinations of space–time covariance functions and variograms. *IEEE Transactions on Signal Processing*, v. 53, n.3, mar, 2005.

39. MINGOTI, S. A. As funções de madograma e rodograma como alternativas para descrever a variabilidade espacial dos dados. *Revista Escola de Minas*, v.49, n. 02, p. 71-74, 1996.

40. MINGOTI, S. A.; LEITE, A. F.; ROSA, G. Describing the total number of diagnosed cases of aids by means of geostatistics. *Rev. Mat. Estat*., São Paulo, v. 24, n.1, p. 61-76, 2006.

41. MINGOTI, S. A.; ROSA, G. A note on robust and non-robust variogram estimators. *Revista Escola de Minas*, v. 61, n.1, p. 87-95, 2008.

42. MORETI, D. *Avaliação espaço-temporal de processos do balanço de água em um solo com citros*. 138p. Tese (Doutorado em Agronomia) – Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2006.

43. NIU, XU-F.; MCKEAGUE, I. W.; ELSNER, J. B. Seasonal space-time models for climate systems. *Statistical Inference for Stochastic Processes*, Netherlands, v. 6, p. 11-133, 2003.

44. PAEZ, M. S. *Análise de modelos para a estimação e previsão de processos espaço-temporais*. 113p. Tese (Doutorado em Estatística) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2004.

45. PAEZ, M. S.; GAMERMAN, D. Modelagem de processos espaço-temporais. In: ESCOLA DE SÉRIES TEMPORAIS E ECONOMETRIA, 11, 2005, Vila Velha, *Minicurso*...São Paulo: Associação Brasileira de Estatística. 102p.

46. PANTUZZO, A. E.; MINGOTI, A. S. Predição do número total de casos diagnosticados de aids nos municípios de minas gerais através de técnicas de estatística espacial. *Rev. Mat. Estat*., São Paulo, v.16, p. 59-80, 1998.

47. PORCU, E.; MATEU, J.; BEVILACQUA, M. Covariance functions that are stationary or nonstationary in space and stationary in time. *Statistica Neerlandica*, v. 61, n. 3, p. 358-382, 2007.

48. R: A language and environment for statistical computing, R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2006, ISBN 3-900051-07-0, http://www.R-project.org.

49. RandomFields: Simulation and analysis of random fields, SCHLATHER, M., R package version 1.3.30, 2006 http://www.stochastik.math.uni-goettingen.de/institute.

50. ROSA, G. *Avaliando a qualidade dos estimadores de variograma (variograma experimental) e do método de mínimos quadrados ponderados para estimação dos parâmetros do modelo de variograma teórico do processo*. 122p. Dissertação (Mestrado em Estatística) – Universidade Federal de Minas Gerais, Belo Horizonte, 2003.

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

52. SCHLATHER, M., Simulation and analysis of random fields. *R News*, Austria, v. 1, p. 18-20, 2001.

53. SCHMIDT, A. M.; SANSÓ, B. Modelagem bayesiana da estrutura de covariância de processos espaciais e espaço-temporais. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 17, 2006, Caxambu, *Minicurso*... São Paulo: Associação Brasileira de Estatística. 151p.

54. SILVA, A. S. D. *Modelos gaussianos geoestatísticos espaço-temporais e aplicações*. 69p. Dissertação (Mestrado em Agronomia) – Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2006.

55. SILVA, A. S. D.; RIBEIRO JÚNIOR, P. J.; ELMATZOGLOU, I. Modelagem geoestatística utilizando a família de gneiting de funções de covariância espaço-temporais. Rev. Mat. Estat., v. 25, n. 1, p. 65-83, 2007.

56. STEIN, M. L. Space time–covariance functions. *Journal of the American Statistical Association*, v. 100, n. 469, p. 310-321, mar, 2005.

57. WALLER, L. A.; GOTWAY, C. A. *Applied spatial statistics for public health data*. New Jersey: John Wiley & Sons, 2004. 494p.