Presidente da República
Luíz Inácio Lula da Silva

Ministro do Planejamento, Orçamento e Gestão
Guido Mantega

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Eduardo Pereira Nunes

Diretor Executivo
José Sant´Anna Bevilaqua

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
Wasmália Socorro Barata Bivar

Diretoria de Geociências
Guido Gelli

Diretoria de Informática
Luiz Fernando Pinto Mariano

Centro de Documentação e Disseminação de Informações
David Wu Tai

Escola Nacional de Ciências Estatísticas
Pedro Luis do Nascimento Silva

Ministério do Planejamento, Orçamento e Gestão
Instituto Brasileiro de Geografia e Estatística - IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 63 número 220 julho/dezembro 2002

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 63, n. 220, p. 1-90, jul./dez. 2002

Instituto Brasileiro de Geografia e Estatística - IBGE
Av. Franklin Roosevelt, 166 - Centro - 20021-120 - Rio de Janeiro - RJ - Brasil

© IBGE. 2004

Revista Brasileira de Estatística, ISSN 0034-7175

Órgão oficial do IBGE e da Associação Brasileira de Estatística – ABE.

Publicação semestral que se destina a promover e ampliar o uso de métodos estatísticos (quantitativos) na área das ciências econômicas e sociais, através de divulgação de artigos inéditos.

Temas abordando aspectos do desenvolvimento metodológico serão aceitos, desde que relevantes para os órgãos produtores de estatísticas.

Os originais para publicação deverão ser submetidos em três vias (que não serão devolvidas) para:

Renato Martins Assunção
Editor responsável – RBEs – IBGE.
Av. República do Chile, 500 – Centro
20031-170 – Rio de Janeiro, RJ.

Os artigos submetidos às RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.

A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Editor Responsável
Renato Martins Assunção (UFMG)

Editor de Estatísticas Oficiais
Pedro Luis do Nascimento Silva (ENCE)

Editor de Metodologia
Francisco Louzada-Neto (UFSCAR)

Editores Associados
Gilberto Alvarenga Paula (USP)
Kaizô Iwakami Beltrão (IBGE)
Djalma Galvão Carneiro Pessoa
Helio dos Santos Migon
Lisbeth Kaiserlian Cordani (USP)
Wilton de Oliveira Bussab (FGV-SP)
Francisco Cribari-Neto

Editoração
Helem Ortega da Silva - Coordenação de Métodos e Qualidade - DPE/COMEQ

Impressão
GráficaDigital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2004.

Capa
Renato J. Aguiar – Coordenação de Marketing/CDDI

Ilustração da Capa
Marcos Balster – Coordenação de Marketing/CDDI

Revista brasileira de estatística/IBGE, - v.1, n.1 (jan./mar.1940), - Rio de Janeiro:IBGE, 1940-

v.

Trimestral (1940-1986), semestral (1987-).

Continuação de: Revista de economia e estatística.

Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v. 50 (1980-1989).

Co-edição com a Associação Brasileira de Estatística a partir do v.58.

ISSN 0034-7175 = Revista brasileira de estatística.

I. Estatística – Periódicos. I. IBGE. II. Associação Brasileira de Estatística.

IBGE. CDDI. Div. de Biblioteca e Acervos Especiais CDU 31 (05)
RJ-IBGE/88-05 (rev.98) PERIÓDICO

Impresso no Brasil/*Printed in Brazil*

Sumário

Nota do Editor

Mais um número da Revista Brasileira de Estatística é publicado este ano. Como é comum em nossa revista, esse número conta com uma grande diversidade de assuntos. Além disso, temos também uma grande diversidade geográfica. O primeiro artigo vem de pesquisadores ligados à Universidade Federal de Minas Gerais e trata de um método de análise de conglomerados espaciais. O segundo artigo conta coma equipe de pesquisadores da Universidade do Tocantins e da Universidade Federal de Viçosa e trata do uso da distância de Mahalanobis na escolha de progenitores em programa de melhoramento de soja para consumo humano. Dado a importância do agronegócio na economia brasileira é ótimo ver os estatísticos participando ativamente dessa área. O terceiro artigo é mais técnico e de um tipo menos comum em nossa revista que possui um caráter mais aplicado. Esse artigo foi escrito por professores da Universidade Federal de Pernambuco e estuda a consistência de estimadores da matriz de covariância sob heterocedasticidade. O último artigo desse número é uma excelente revisão de modelos para dados espacialmente referenciados e foi escrito por pesquisadores da Universidade Federal do Rio de Janeiro. Esta é uma área de crescente interesse entre usuários que geram um número cada vez maior de bases de dados georeferenciadas. Além disso, softwares baratos ou até mesmo gratuitos (ver, por exemplo, o software TERRACRIME) encontram-se disponíveis para análises. Apesar da grande sofisticação técnica existente no momento, o desenvolvimento de técnicas estatísticas para este tipo de dados ainda está na sua infância e resta muito a ser feito.

Estamos às vésperas de mais um SINAPE e a RBEs espera ver muitas submissões vindas de artigos submetidos ao congresso. Se isto ocorrer temos uma chance de colocar a revista em dia e assim podermos passar a lutar pela sua expansão do numero de assinantes.

Boa leitura de mais um número da Revista Brasileira de Estatística.

Renato Martins Assunção
Editor Responsável

Análise de Conglomerados Espaciais via Árvore Geradora Mínima

Renato M. Assunção*
Juliano P. Lage**
Edna A. Reis*

Resumo

Em Econometria, Epidemiologia e Geografia é freqüente a necessidade de regionalizar um mapa. Isto é, partindo de uma região dividida em pequenas áreas, queremos agregá-las em alguns poucos conglomerados de áreas contíguas. Esses conglomerados devem ser compostos internamente de áreas muito similares e, ao mesmo tempo, conglomerados distintos devem ser formados por áreas muito dissimilares. A similaridade diz respeito a um vetor de variáveis medido em cada uma das áreas. Este problema é idêntico ao problema de análise de conglomerados usual a não ser pela restrição de contiguidade das áreas. Apresentamos uma proposta para criar conglomerados espaciais através da partição sucessiva da árvore geradora mínima associada ao grafo da região. Implementamos a metodologia no software SKATER e ilustramos sua aplicação com dados de condições de vida nos municípios de Minas Gerais.

* Endereço para correspondência: Departamento de Estatística/ICEx/UFMG, CEP - 31270-901, Belo Horizonte, MG. E-mail: assuncao@est.ufmg.br.
** Departamento de Computação /ICEx/UFMG, CEP - 31270-901, Belo Horizonte, MG.

1. Introdução

Os métodos usuais de Análise de Conglomerados abordam o problema de reunir objetos em grupos internamente homogêneos em relação a um conjunto de características medidas em cada objeto. Ao mesmo tempo, objetos de diferentes grupos devem ser heterogêneos entre si. Em Estatística Espacial, o problema reside em realizar uma análise de conglomerados quando esses objetos possuem uma localização espacial. A situação é a de uma região dividida em pequenas áreas, cada uma delas com uma posição geográfica determinada, como os setores censitários que compõem um município ou os municípios que formam um estado. Para cada área, tem-se as medidas de um conjunto de variáveis de interesse, tais como suas características sociais e econômicas, que formam o perfil da área. Deseja-se reunir essas pequenas áreas em regiões disjuntas que atendam a duas condições simultaneamente: áreas de uma mesma região devem ser similares com relação às variáveis do perfil e, ao mesmo tempo, áreas de regiões diferentes devem ser dissimilares. A grande diferença em relação à análise de conglomerados usual é a necessidade de que essas regiões sejam formadas por áreas contíguas no espaço. Este procedimento é chamado de *regionalização* em Economia Regional e Geografia.

A análise de conglomerados usual não garante que a partição final será formada por conglomerados espaciais, a não ser através da ação subjetiva e manual do usuário. Será comum termos áreas geograficamente distantes agrupadas no mesmo conglomerado pela análise usual.

Para realizar uma regionalização, pode-se fazer uso de especialistas que, sendo conhecedores da região, procedem a uma subdivisão sem o uso de critérios objetivos. A desvantagem óbvia deste procedimento é sua grande subjetividade. A dificuldade de se reproduzir os resultados obtidos e de avaliar os critérios de fato utilizados na subdivisão são as grandes desvantagens desse procedimento.

Assim como na análise de conglomerados usual, para encontrar a melhor divisão das áreas em p conglomerados, seria necessário listar todos os possíveis conglomerados de tamanho p e, dentre eles, buscar aquele que maximize a homogeneidade dentro dos conglomerados e a heterogeneidade entre eles. Este procedimento seria repetido para cada valor possível de p. Finalmente, dentre todas as partições ótimas obtidas em cada valor de p, seria escolhida aquela que resultasse na melhor divisão, segundo algum critério.

Entretanto, existe uma enorme dificuldade computacional com esse procedimento de minimização sobre todas as possíveis regionalizações. Mesmo com um super computador, o procedimento levaria dias para ser executado se o número de áreas for moderado (algumas

poucas centenas de áreas, por exemplo). Assim, em essência, todos os métodos de análise de conglomerados procuram construir uma classe menor de possíveis partições que devem ser pesquisadas na busca do mínimo de uma função-objetivo. Essa classe de partições deve atender a dois requisitos: ela deve ser pequena o suficiente para que possa ser pesquisada rapidamente e, ao mesmo tempo, ela deve ser rica o suficiente para que o ótimo global seja próximo do ótimo encontrado dentro da classe menor.

A dificuldade de se listar todas as possíveis divisões das áreas em p regiões encontradas na análise de conglomerados usual ganha mais uma restrição em estatística espacial: a de que as áreas de cada região sejam contíguas. Uma proposta para este problema foi feita por Carvalho et al (1996) buscando adaptar os procedimentos de análise de conglomerados usuais. Haining (1996) também usou um procedimento não-espacial baseado em teoria da informação para formar conglomerados. Um procedimento que respeita a restrição de contiguidade espacial foi proposto por Carvalho et al (1998). Osnes (1999) usou índices de autocorrelação espacial para agregar iteradamente áreas contíguas formando, assim, conglomerados espaciais. Rezende et al. (2000) usaram *diagramas de Voronoi* para definir áreas de interesse da saúde pública no Município do Rio de Janeiro. Kiang (2001) usou técnicas de redes neurais para criar conglomerados espaciais. Maravalle e Simeone (1995) propuseram uma abordagem baseada em grafos de contiguidade espacial e sua substituição por árvores geradoras. Num trabalho posterior, eles estudaram a complexidade computacional desses métodos (Maravalle et al., 1997).

Não existem princípios teóricos gerais a partir dos quais sejam derivados procedimentos ótimos de criação de conglomerados, sendo usados apenas argumentos heurísticos. Como conseqüência, uma enorme variedade de métodos surgiram para a análise não-espacial de conglomerados e nenhum é claramente dominante em relação aos demais. Esses métodos são *ad hoc* e diferem nos critérios de construir os grupamentos e nas formas de medir a qualidade dos resultados obtidos. Infelizmente, não existe ainda uma forma objetiva e padronizada para avaliar o desempenho dos diversos procedimentos existentes.

Neste trabalho, seguimos o trabalho inicial de Maravalle e Simeone (1995) transformando o mapa num grafo e reduzindo-o a uma árvore geradora. Nós escolhemos a árvore geradora mínima para fazer esta redução. A partir daí, particionamos, sucessivamente, a árvore geradora para obter a regionalização. Desta forma, nosso método usa uma classe de partições espaciais que procura atender aos dois requisitos acima: ser grande o bastante para conter soluções satisfatórias mas não ser tão grande que inviabilize uma busca algorítmica do seu ótimo. Diferentemente de Maravalle e Simeone (1995), não otimizamos localmente a árvore geradora mínima com respeito a uma função-objetivo de partição em conlglomerados, pois o ganho

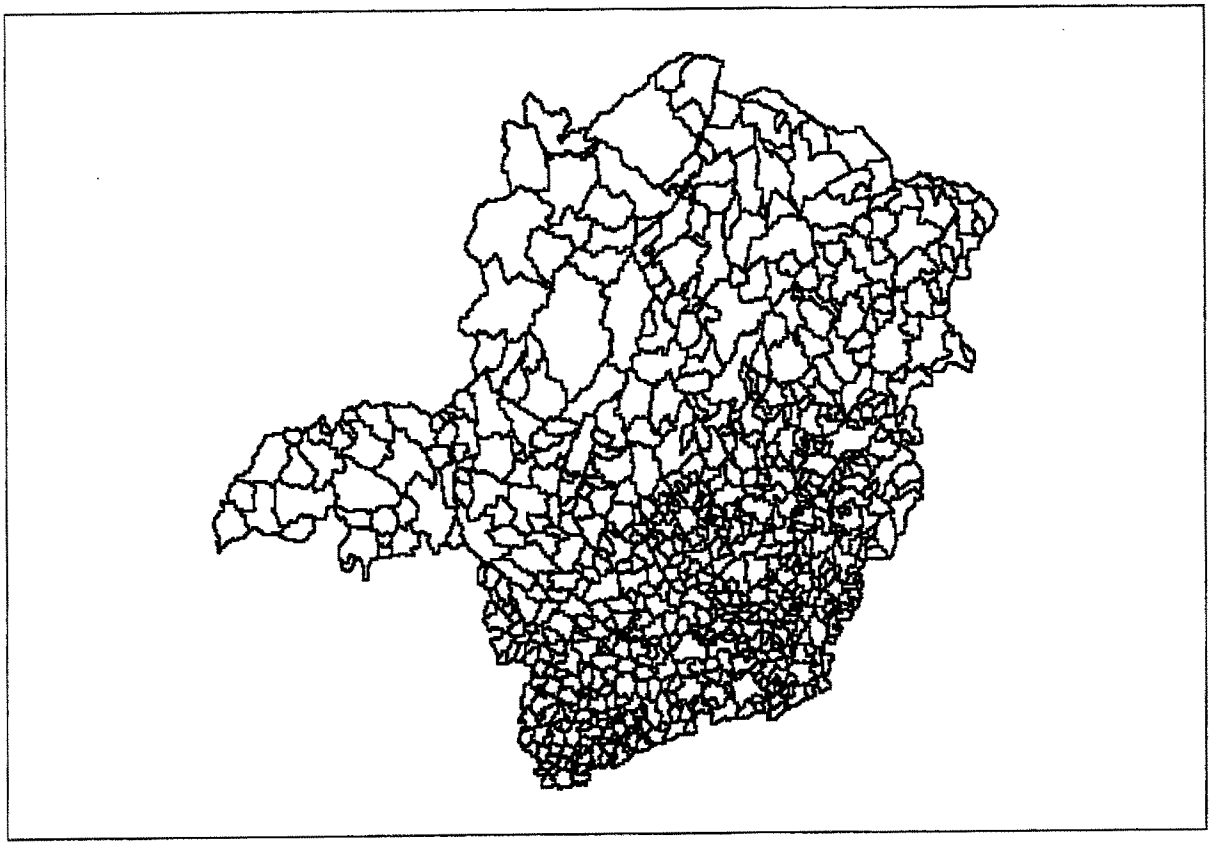
incremental, em nossa opinião, será irrisório e computacionalmente intensivo. Solicitamos o programa desenvolvido por esses autores para compararmos com nosso método, mas não obtivemos resposta. Assim, não pudemos fazer uma comparação entre as duas diferentes abordagens.

Nosso método foi implementado no software SKATER, que possui uma interface gráfica que o torna de fácil uso, especialmente a entrada e saída de dados. Ilustramos a metodologia através da aplicação dos dados sobre condições de vida para os municípios de Minas Gerais.

2. Conglomerados espaciais

Vamos considerar n áreas geográficas contíguas organizadas na forma de um mapa, como os 723 municípios de Minas Gerais (Figura 1). Associado a cada área i, $i = 1, 2,.., n$, temos um vetor $x_i = (x_{i1}, x_{i2},...,x_{im})$ de m características quantitativas constituindo o perfil dessa área.

Figura 1 - Mapa do Estado de Minas Gerais dividido em 723 municípios



Para estudar os métodos de análise espacial, vamos introduzir algumas definições. Duas áreas são consideradas *vizinhas* quando elas possuem uma fronteira comum.

Um *conglomerado* é qualquer subconjunto de áreas. Conglomerados serão interessantes apenas se eles forem internamente homogêneos. Entretanto, esta característica não faz parte da definição de conglomerado, ela é apenas uma propriedade desejável a ser perseguida. Conglomerados podem ser constituídos por uma única área.

Um conglomerado é *conectado* se, para toda subdivisão do conglomerado em dois subconjuntos disjuntos e complementares, pelo menos uma área de um subconjunto é vizinha de pelo menos uma área do outro subconjunto. Conglomerados formados por uma única área são considerados conectados. Assim, em um conglomerado conectado, qualquer área possui alguma vizinha no mesmo conglomerado.

Uma região está particionada em *conglomerados espaciais* quando as áreas que formam esta região estiverem agrupadas em conglomerados que sejam disjuntos e conectados.

3. O método da árvore geradora mínima

Nesta abordagem, representamos o mapa das n áreas por um *grafo* onde cada área i corresponde a um *nó* v_i ($i = 1,\ldots,n$) e duas áreas vizinhas i e j são ligadas por uma *aresta* (v_i, v_j). A Figura 2 mostra à esquerda um grafo com dez nós localizados nos centróides das áreas de um mapa. A existência de uma aresta indica que as áreas são vizinhas.

Um *caminho* de v_i a v_k é uma seqüência de nós distintos $v_i, v_j, \ldots, v_k$ que são conectados pelas arestas $(v_i, v_j), (v_j, v_l), \ldots, (v_m, v_k)$.

Um grafo é dito *conexo* se, para ir de um nó v_i do grafo a qualquer outro nó v_j, existe pelo menos um caminho de v_i a v_j. Usando nossa definição de vizinhança, vamos sempre encontrar o grafo correspondente ao mapa como um grafo conexo se o mapa não apresentar ilhas.

A cada aresta, vamos associar um *custo* relacionado ao grau de dissimilaridade dos perfis entre duas áreas. Quanto maior a dissimilaridade entre duas áreas maior será este custo. Existem várias alternativas possíveis para definir o custo de uma aresta. Por exemplo, o custo da aresta que une as áreas i e k pode ser a distância euclidiana entre as variáveis do perfil:

$$Custo(i,k) = \sqrt{\sum_{j=1}^{m} (x_{ij} - x_{kj})^2} \quad . \qquad (1)$$

É comum que as variáveis estejam padronizadas de alguma forma antes de calcular o custo pois, caso contrário, as variáveis com maior variância tendem a dominar o valor da dissimilaridade. Outra opção possível é a distância de Mahalanobis (Seber, 1984). No grafo à esquerda na Figura 2, a espessura da linha é proporcional ao custo daquela aresta. Isto é, a espessura da aresta é proporcional à dissimilaridade das duas áreas medida como a distância entre os vetores de seus atributos.

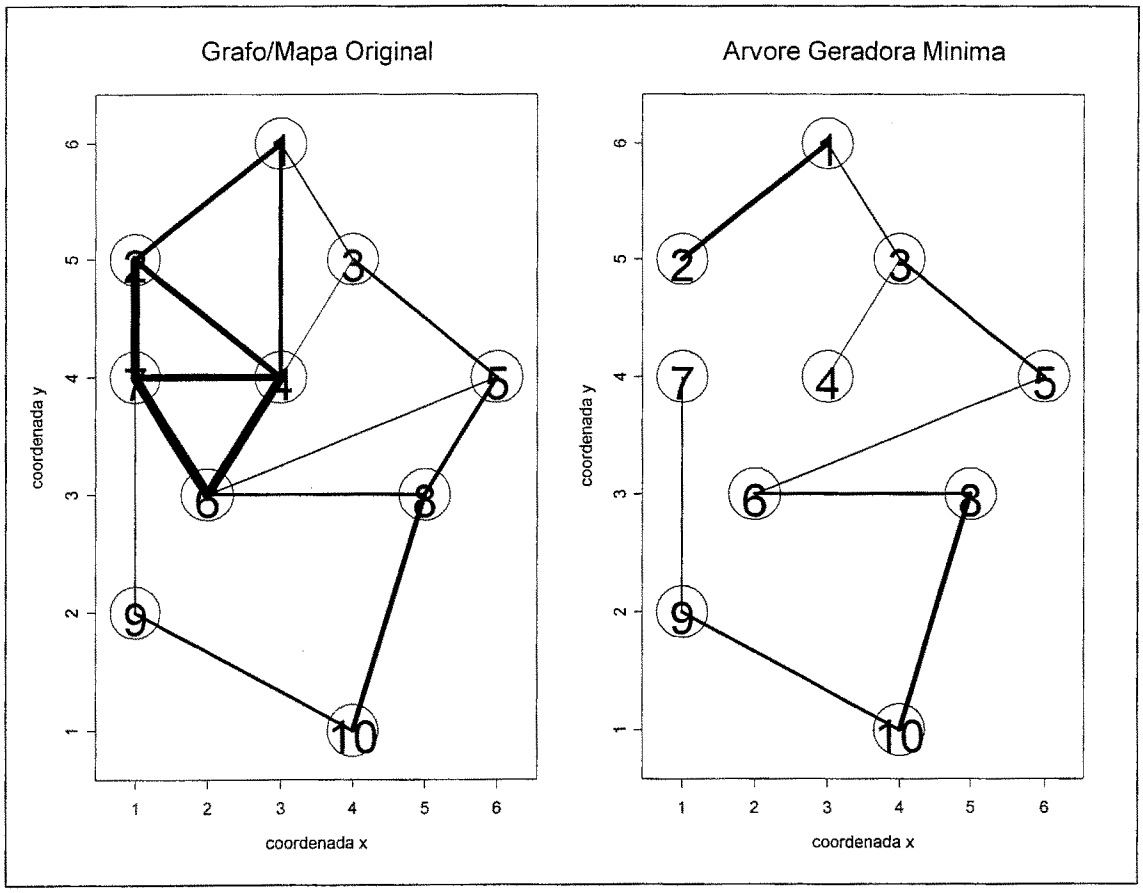
A idéia inicial do nosso procedimento é simplificar o grafo apagando arestas de forma a ficar com um grafo reduzido, mas ainda conexo. Isto é, deverá ainda ser possível sair de uma área e chegar a qualquer outra área do mapa percorrendo, sucessivamente, arestas do grafo. Mas queremos apagar, principalmente, aquelas arestas de custo mais elevado de modo que, ao saltar de uma área a outra, tenhamos uma diferença pequena nos atributos das áreas envolvidas no salto. Além disso, queremos terminar com uma árvore tal que, se apagarmos qualquer aresta adicional, o grafo ficará dividido em dois subgrafos desconectados, os quais serão os candidatos a constituírem dois conglomerados espaciais. Isto é possível através da construção de uma *Árvore Geradora Mínima* do grafo, como explicamos a seguir.

Vamos introduzir mais alguns conceitos básicos necessários. Um *circuito* num grafo é um caminho onde o nó inicial e o nó final são os mesmos. Uma *árvore* é um grafo conexo que não contém circuitos. Uma *árvore geradora* para um grafo G é um subgrafo que é uma árvore e que contém todos os nós de G. Assim, em uma árvore geradora, quaisquer dois nós são unidos por um único caminho. Além disso, o número de arestas é igual ao número de nós menos um. Isso implica que, se qualquer aresta é apagada, a árvore estará desmembrada em duas subárvores desconectadas. O *custo de um grafo* é a soma dos custos das arestas do grafo. Uma *árvore geradora mínima* é uma árvore geradora que possui custo mínimo.

No gráfico do lado direito na Figura 2, encontra-se a árvore geradora mínima do grafo localizado na figura da esquerda. Note que, se apagarmos qualquer aresta na árvore da direita, teremos dois subgrafos desconectados.

Um grafo pode ter mais que uma árvore geradora mínima, especialmente se os custos assumirem valores apenas em um conjunto discreto finito de valores possíveis. Porém, dificilmente esse será o caso nas situações a serem consideradas na prática. O motivo é que o custo de uma aresta será uma medida de distância entre dois vetores de atributos, um para cada uma das duas áreas componentes da aresta. Como essa distância será, geralmente, a distância euclidiana ou a distância de Mahalanobis e pelo menos algumas das variáveis serão contínuas, dificilmente teremos duas arestas com custos idênticos. Com probabilidade 1, a árvore geradora mínima é única se pelo menos uma das variáveis for contínua e uma dessas distâncias.

Figura 2 - Grafo de mapa de região hipotética dividida em dez áreas e a árvore geradora mínima correspondente. As arestas que ligam um par de área têm espessura proporcional a seu *custo*



3.1. Algoritmo de Prim para árvore geradora mínima

Utilizamos o algoritmo de Prim (1957) para construir uma árvore geradora mínima. A partir de um grafo conexo com custos associadas às arestas, construímos a árvore de forma recursiva, começando com a árvore T_1 e aumentado-a progressivamente T_2, T_3, ..., até T_n, que é a árvore geradora mínima.

Passo 1: Tome qualquer nó v e faça $T_1 = v$.

Repita o passo 2 tanto quanto possível:

Passo 2: Dentre as arestas que unem um nó em T_k a um nó que não está em T_k., encontre aquela aresta e_k de menor custo. Se existir mais de uma aresta com essa propriedade, escolha uma delas arbitrariamente. A árvore T_{k+1} é o grafo obtido adicionando essa aresta e_k e seu nó terminal à T_k.

Como o número de nós é finito, o algoritmo é interrompido em algum momento. Nenhuma aresta da árvore geradora mínima pode ser substituída sob pena de gerar uma árvore de custo maior ou igual. Apesar disso, não é verdade que o caminho entre quaisquer duas áreas na árvore geradora mínima seja aquele de custo mínimo entre estas duas áreas no grafo original. O custo mínimo da árvore geradora mínima é uma propriedade global e não uma propriedade local

3.2. Critérios de poda da árvore geradora mínima

Após a criação da árvore geradora mínima, passamos a particioná-la para obter os conglomerados espaciais. O problema combinatório de formação dos conglomerados espaciais está agora bastante reduzido, pois basta verificar as n arestas da árvore e apagar uma delas para ter a árvore separada em dois subgrafos desconectados. Iterando este procedimento que apaga arestas em cada subgrafo resultante, os conglomerados vão sendo criados de forma hierárquica.

Dada uma árvore geradora mínima, uma escolha natural da aresta a ser apagada é aquela que possui o maior custo ou dissimilaridade. Ao apagar esta aresta, teremos como resultado dois subgrafos desconectados, cada um deles conexo, que podem ser vistos como dois conglomerados espaciais. O custo desse novo grafo bipartido é a soma dos custos das arestas não apagadas nos dois subgrafos. É claro que, se qualquer outra aresta fosse apagada na árvore inicial, o resultado seria um grafo dividido em dois com um custo maior (ou igual). Nesse sentido, é natural escolhermos para apagar a aresta de custo máximo. Continua-se dividindo a árvore apagando-se, sucessivamente, a aresta com o segundo menor custo (criando três subgrafos desconectados), apagando a aresta com o terceiro menor custo (criando quatro subgrafos), etc.

No entanto, a escolha da aresta a ser apagada usando a mesma medida de dissimilaridade usada para construir a árvore geradora mínima tem duas grandes desvantagens. Ocorre que as últimas arestas a serem adicionadas na árvore geradora mínima tendem a ter os maiores custos. Afinal, não é sem motivos que essas áreas foram as últimas a serem conectadas na árvore geradora mínima. Ao apagar as arestas de maior custo na árvore geradora mínima, estaremos tendendo a quebrar o grafo nas últimas arestas que foram adicionadas à árvore as quais, por sua vez, tendem a ligar áreas isoladas no grafo. Isto é, apagar as últimas arestas tende a gerar dois conglomerados, um formado por apenas umas poucas áreas e o outro, com o restante das áreas. A outra desvantagem é que o custo de uma aresta é uma medida de dissimilaridade entre duas áreas e, portanto, de caráter local. Como o objetivo é formar conglomerados homogêneos internamente, nosso critério de partição da árvore deveria respeitar este objetivo.

Assim, buscamos uma definição alternativa de custo para esta segunda etapa onde arestas são sucessivamente apagadas do grafo da árvore geradora mínima. Numa árvore, denotamos por SSTO a soma de quadrados dos desvios no espaço das variáveis em relação à média de todas as áreas da árvore, ou seja:

$$SSTO = \sum_{j=1}^{m} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2 \qquad .. \qquad (2)$$

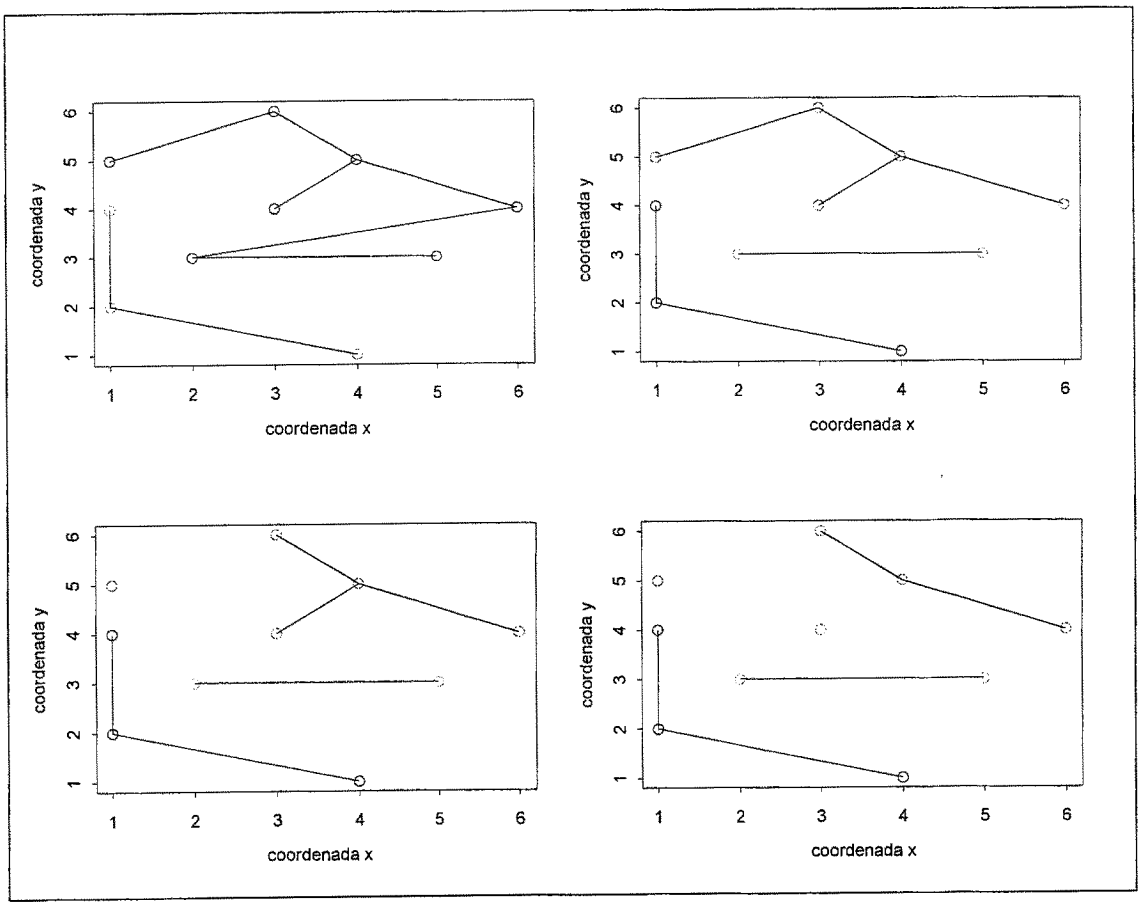
onde $\bar{x}_j = \sum_{i=1}^{m} \frac{x_{ij}}{n}$.

Para cada um dos dois conglomerados resultantes após apagada uma aresta na árvore geradora mínima, calculamos a soma de quadrados dos desvios no espaço das variáveis em relação à média do conglomerado resultante. A seguir, somamos as duas somas de quadrados, uma de cada conglomerado resultante, obtendo SSA. Quanto menor SSA, mais homogêneos serão os conglomerados resultantes. Como SSA está entre 0 e SSTO, definimos o *custo de apagar a aresta* como sendo SSTO-SSA. Assim, um custo alto associado à aresta indica que seu desaparecimento vai gerar conglomerados homogêneos.

Escolhemos apagar a aresta de maior *custo de apagar*, gerando, assim, os dois primeiros conglomerados. Este procedimento será repetido dentro de cada conglomerado resultante, mas agora vamos apagar a aresta de maior *custo de apagar* entre aquelas dos dois conglomerados. Este processo termina quando todas as arestas forem apagadas, gerando n conglomerados, cada um composto de uma das n áreas.

A Figura 3 a seguir mostra o resultado de adotarmos esse novo critério para particionar a árvore geradora mínima do exemplo simples com dez áreas da Figura 2.

Figura 3 - Partições sucessivas da árvore geradora mínima do exemplo de dez áreas pelo critério da maximização da redução da soma de quadrados dos conglomerados resultantes



Para escolher o número de conglomerados na regionalização final, ou seja, definir quando parar a poda da árvore geradora mínima, sugerimos analisar o gráfico da SSTO-SSA *versus* o número de *clusters*, em cada estágio de poda. Esta curva representa uma função monótona não crescente e espera-se que, nas primeiras partições, ela diminua substancialmente diminuindo o ritmo de decrescimento à medida que um grande número de conglomerados já estiverem presentes.

O critério de parada pode também ser definido pelo usuário. Um critério poderia ser, por exemplo, podar a árvore geradora mínima até que a população dos conglomerados resultantes não seja menor que um valor determinado.

4. O software SKATER

Inicialmente, o método de análise de conglomerados espaciais via partição da árvore geradora mínima apresentado neste trabalho foi implementado computacionalmente através de um conjunto de funções da linguagem *R*. Entretanto, as funções *R* são lentas quando comparadas ao que é possível obter com linguagens compiladas, tais como C++ ou Delphi. Além disso, programas compilados nessas linguagens não dependem da existência de um software tal como o *R*. Desse modo, criamos o software SKATER (*Spatial 'K'luster Analysis by Tree Edge Removal*), uma ferramenta amigável em linguagem C++ para ambiente Windows®.

A estrutura dos dados de entrada no SKATER é muito simples, constituída por dois arquivos do tipo texto. Não é necessário ter o mapa digitalizado da região como o da Figura 1. Para definir a localização espacial das áreas (ou nós do grafo), são necessárias apenas as coordenadas do centróide de cada área. A informação de vizinhança é fornecida em arquivo no qual lista-se cada área (representada por um índice) e suas respectivas áreas vizinhas. Em aplicações onde o número de áreas não é muito grande, estas informações podem ser obtidas até mesmo manualmente.

O software desenha o grafo completo da região, o grafo da árvore geradora mínima e outro mostrando os conglomerados escolhidos pelo usuário. A listagem das áreas pertencentes a cada conglomerado também pode ser exportada em um arquivo texto.

O software produz o gráfico da queda dos desvios *versus* o número de conglomerados para auxiliar na escolha do número de conglomerados. Outra opção disponível como critério de parada das partições da árvore geradora mínima é informar o tamanho mínimo da população por conglomerado.

O conjunto de funções em *R* e o SKATER pode ser obtido gratuitamente no *site* do Laboratório de Estatística Espacial – LESTE (Departamento de Estatística, Universidade Federal de Minas Gerais), no endereço *www.est.ufmg.br/leste.*

5. Regionalização dos municípios de Minas Gerais usando índices de condições de vida

Como ilustração, a metodologia de criação de conglomerados espaciais via partição da árvore geradora mínima é aplicada ao problema de regionalizar os municípios de Minas Gerais, segundo os Índices de Condições de Vida – ICV, definidos pela Fundação João Pinheiro e IPEA (1996), com base nas informações do Censo Demográfico 1991.

São quatro variáveis que compõem o perfil de condição de vida em cada município: ICV Sáude, ICV Renda, ICV Educação e ICV Criança. Cada índice é composto por um conjunto de indicadores e tem valores em uma escala contínua entre zero e um, de tal forma que valores mais elevados indicam melhores condições de vida.

O mapa da Figura 1 é representado pelo grafo da Figura 4, produzido no SKATER. Esse grafo deve ser visto apenas como uma representação simples do mapa, já que ele não possui a razão de aspecto cartográfica adequada. O grafo da árvore geradora mínima é mostrado na Figura 5.

Figura 4 - Representação dos municípios de Minas Gerais na forma de grafo, com vizinhança definida pelo compartilhamento de fronteira

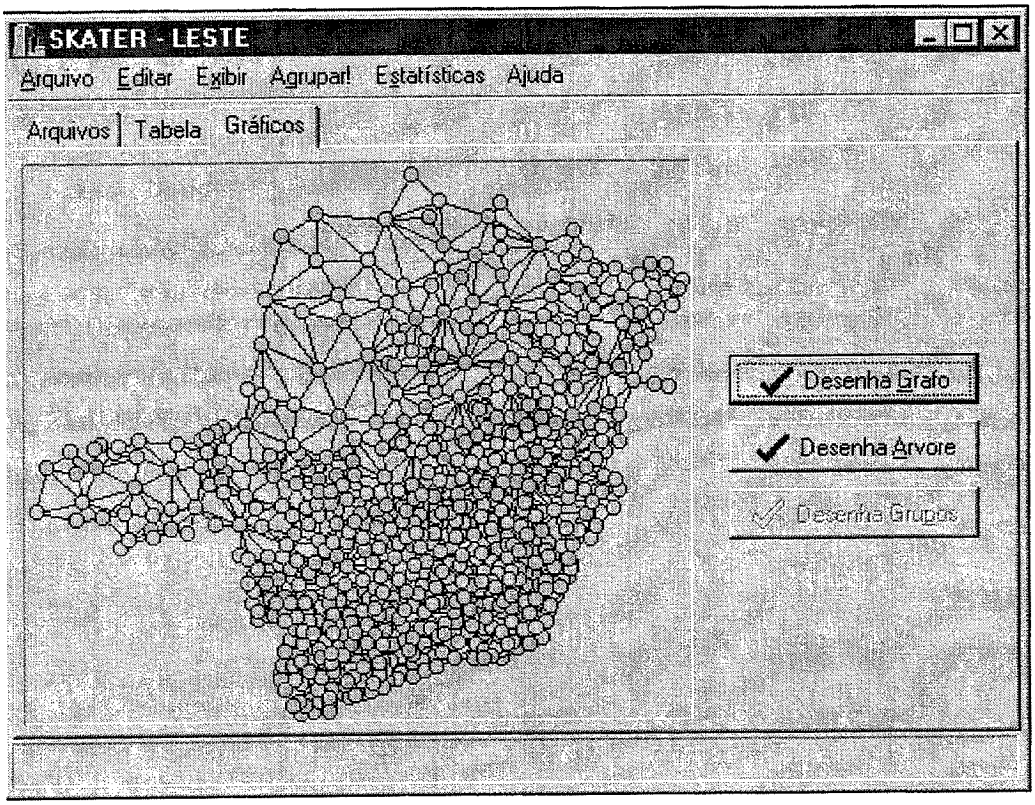
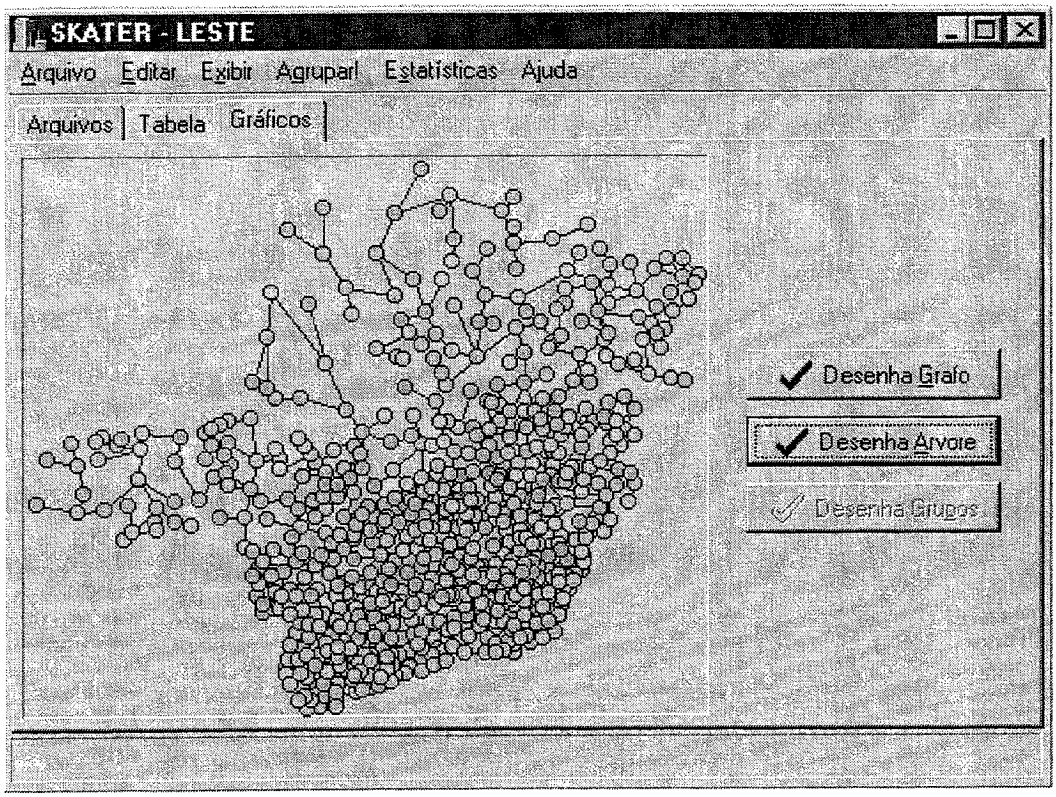
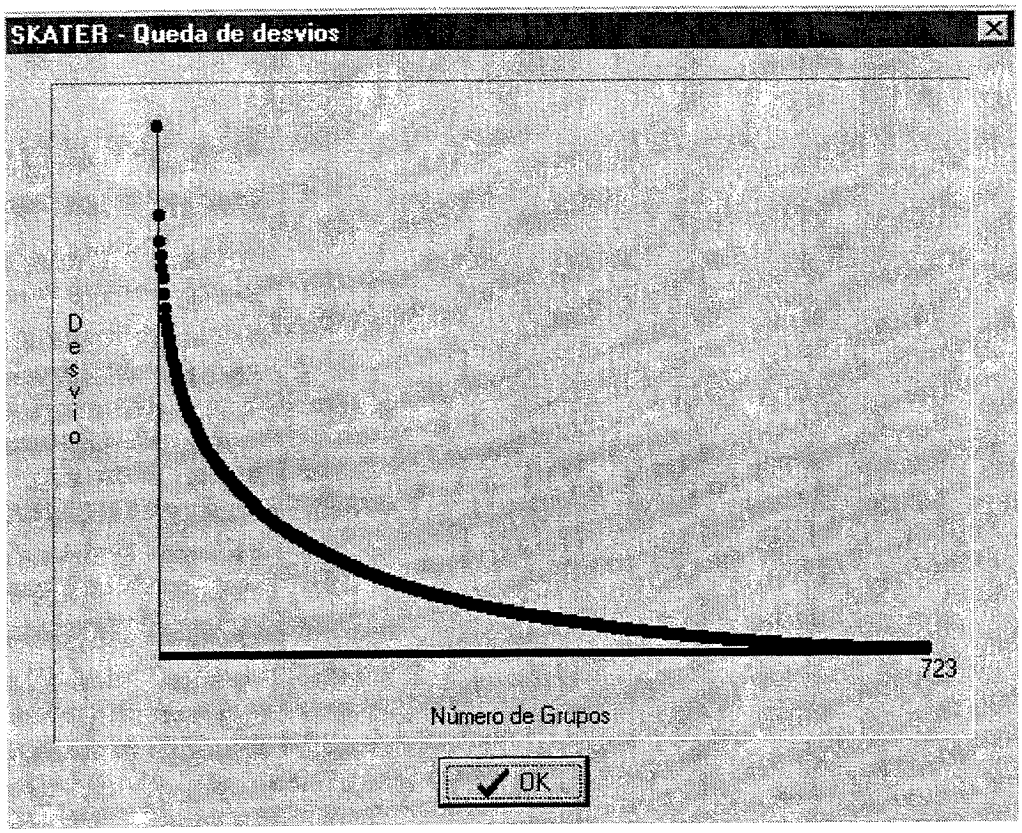


Figura 5 - Árvore geradora mínima dos municípios de Minas Gerais



A redução sucessiva na soma de desvios das variáveis a cada partição da árvore geradora mínima é mostrada na Figura 6.

Figura 6 - Soma de quadrados após apagar, sucessivamente, arestas da árvore geradora mínima dos municípios de Minas Gerais



Para ilustrar a regionalização, a Figura 7 mostra a partição da árvore geradora mínima dos municípios de Minas Gerais em dez conglomerados. Os resultados da partição feita pelo SKATER podem ser exportados para um arquivo texto e, posteriormente, importados num programa de visualização de mapas. O mapa dos municípios de Minas Gerais agrupados nos dez conglomerados é mostrado na Figura 8. Na prática, o número de conglomerados deveria ser maior, mas apresentamos apenas os dez primeiros formados para melhor visualização dos resultados.

Figura 7 - Partição da árvore geradora mínima dos municípios de Minas Gerais em dez conglomerados

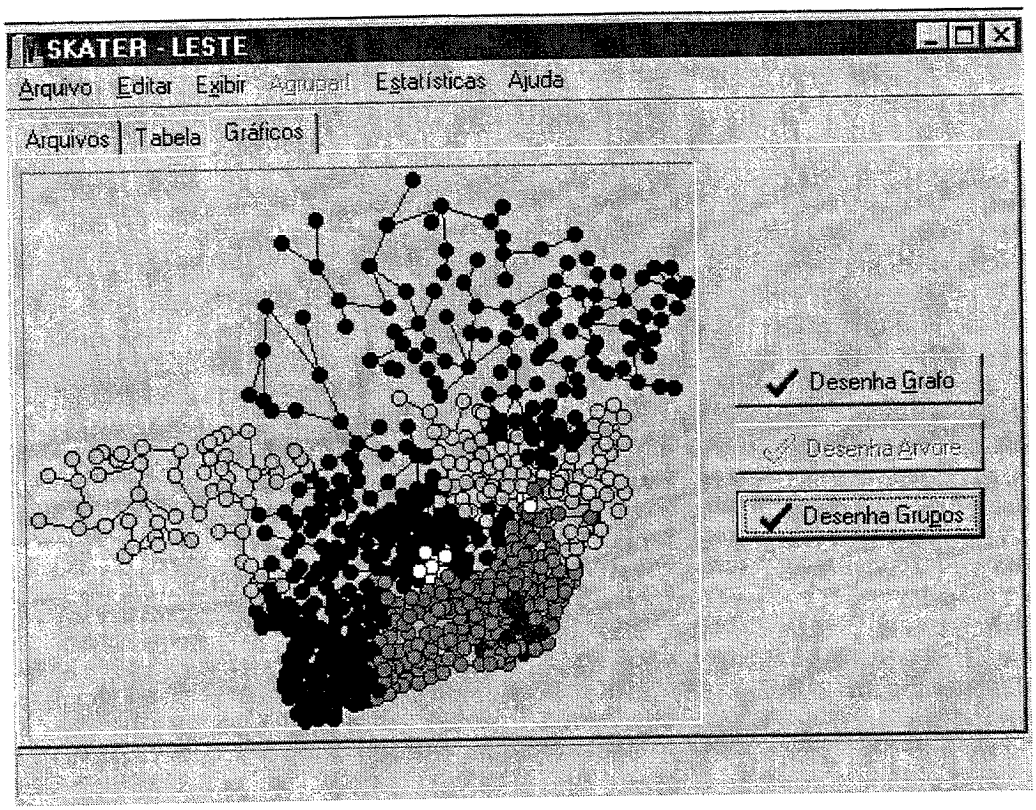
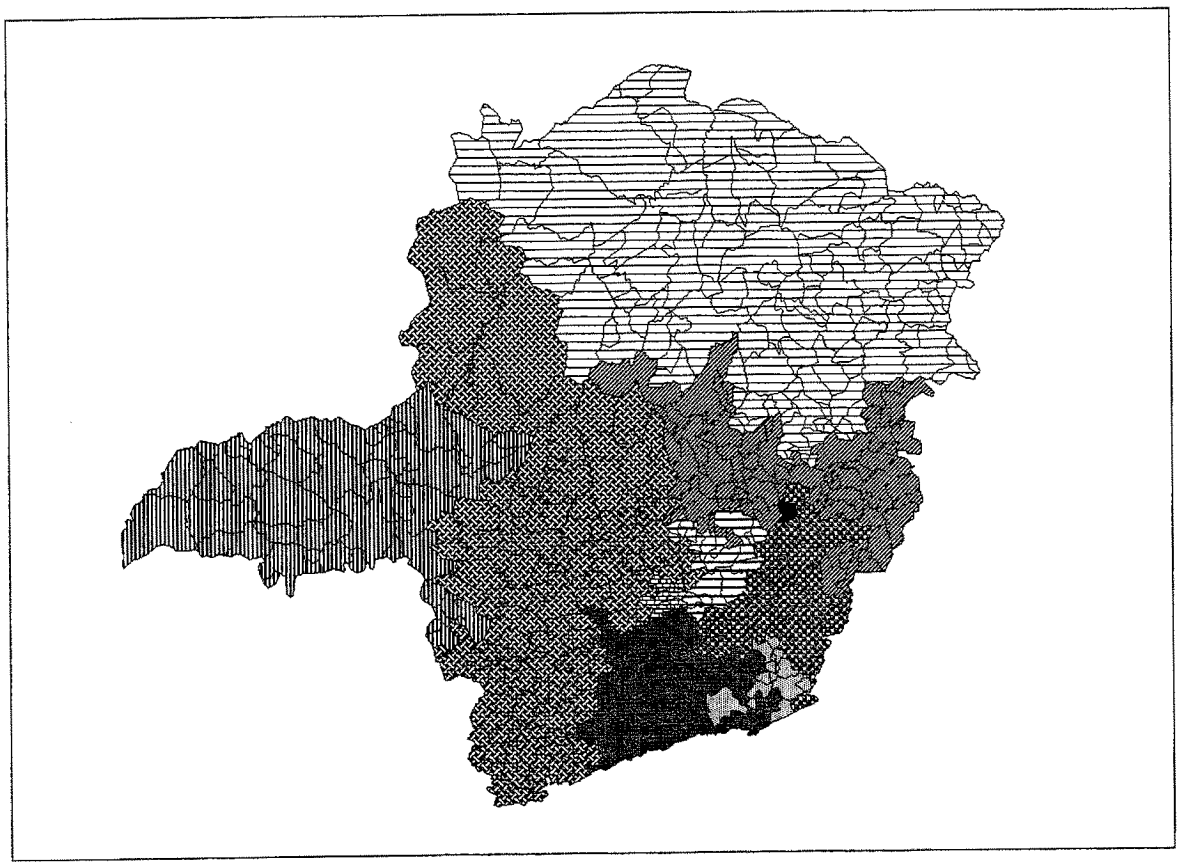


Figura 8 - Mapa dos muncípios de Minas Gerais particionado em dez conglomerados



6. Conclusão

O método de análise de conglomerados espaciais via árvore geradora mínima discutido neste trabalho leva a criação de regiões de áreas contíguas no espaço reduzindo bastante o trabalho que se tem de procura da partição ótima. O uso do método torna-se fácil com sua implementação no software SKATER.

Algumas questões de pesquisa acadêmica foram levantadas ao longo desse trabalho. Uma questão é a possibilidade do método baseado na árvore geradora mínima seguido de sua poda pelo critério de maximização da redução da soma de quadrados dos desvios satisfazer algum critério de otimalidade. Isso poderia dar uma diferente motivação para esse método. Nesse momento, ele é um método heurístico, mas com um critério de otimalidade ficaria mais claro saber em que condições esse método é superior a outros, estabelecendo-se de forma mais precisa seus pontos fortes e fracos.

Outra questão importante é a possibilidade de adaptação do método à situação em que as variáveis medidas nas áreas são taxas calculadas com base nos diferentes tamanhos de população entre as áreas. Neste caso, as estimativas têm graus de precisão diferentes e esta variabilidade deveria ser levada em conta na definição do custo de cada aresta, seja na fase de montagem ou partição da árvore geradora mínima. Dentro desta mesma idéia, seria interessante possibilitar o uso de diferentes pesos para as variáveis do perfil, conforme sua homogeneidade ou grau de importância.

Referências bibliográficas

BREIMAN L, FRIEDMAN J, OLSHEN R, STONE C. (1984) ***Classification and Regression Trees.*** **Wadsworth, Inc.: Pacific Grove, California.**

CARVALHO MS, CRUZ OG, NOBRE FF. (1996) The use of multivariate cluster analysis and aggregation algorithm to propose a new spatial partitioning of Rio de Janeiro, Brazil. ***Statistics in Medicine,*** **15**

CARVALHO MS, CRUZ OG, NOBRE FF. (1998) Fuzzy classification applied to modeling socioeconomic profiles for spatial health analysis. ***Proceedings of the EIS'98 - International ICSC Symposium on Engineering of Intelligent Systems,*** **Tenerife, CD-ROM, paper number 812-170.**

FUNDAÇÃO JOÃO PINHEIRO e INSTITUTO DE PESQUISA ECONÔMICA E APLICADA (1996), ***Condições de Vida nos Municípios de Minas Gerais 1970, 1980 e 1991.*** **Fundação João Pinheiro, Belo Horizonte.**

HAINING R. (1996) Classifying space and analyzing the consequences: spatial analysis of health data. In: ***From Data to knowledge,*** **Gaul W. e Pfeifer D. (eds). Springer-Verlag, Berlin.**

MARAVALLE M, SIMEONE B, NALDINI R. (1997) Clustering on trees. ***Computational Statistics and Data Analysis,*** **24, 217-234.**

OSNES K. (1999) Iterative random aggregation of small units using regional measures of spatial autocorrelation for cluster localization. ***Statistics in Medicine,*** **18, 707-725.**

PRIM RC. (1957) Shortest connection networks and some generalizations, ***Bell Systems Technical Journal,*** **36, 1389.**

REZENDE FS, ALMEIDA RV, NOBRE FF. (2000) Diagramas de Voronoi para a definição de áreas de abrangência de hospitais públicos no município do Rio de Janeiro. ***Cadernos de Saúde Pública,*** **16, 467-475.**

SEBER GAF. (1984) ***Multivariate Observations.*** **Wiley and Sons, New York.**

Agradecimentos

Agradecemos a Pedro Luis do Nascimento Silva, do Departamento de Metodologia do Instituto Brasileiro de Geografia e Estatística - IBGE, por idéias e sugestões ao longo deste trabalho e a Bráulio Figueiredo Alves da Silva, do Centro de Estudos em Criminalidade e Segurança Pública da Universidade Federal de Minas Gerais - CRISP/UFMG, pela valiosa ajuda na produção das figuras.

Abstract

In Econometrics, Epidemiology and Geography, it is frequently necessary to carry out a regionalization of a map. That is, starting from a region partitioned in small areas, we want to aggregate them in few clusters of contiguous areas. Within these clusters we should have similar areas and, at the same time, areas from different clusters should be as dissimilar as possible. Similarity is measured with respect to a vector of attributes defined in each one of the small areas. This problem is identical to a cluster analysis with an added contiguity restriction in the clusters formation. We introduce a proposal to create spatial clusters through the partition of the minimal spanning tree associated to the region graph. We implement our method in the software SKATER and illustrate its application with living condition data in Minas Gerais municipalities.

Distância generalizada de Mahalanobis na escolha de progenitores em programa de melhoramento de soja para consumo humano

Lucas Koshy Naoe*
Newton Deniz Piovesan*
Carlos Sigueyuki Sediyama**
Everaldo Gonçalves de Barros***
Maurílio Alves Moreira****

Resumo

Este trabalho foi realizado com linhagens promissoras de soja oriundas do programa de melhoramento da soja para alimentação humana. As linhagens foram previamente selecionadas para ausência de todas as três lipoxigenases nas sementes e da subunidade protéica A5A4B3. Foram utilizadas oito características agronômicas na indicação de grupos distintos para cruzamento, com base na metodologia de agrupamento de Tocher, utilizando a divergência genética entre as linhagens, estimada por meio da técnica multivariada da distância generalizada de Mahalanobis. Com base nos resultados obtidos, conclui-se que há diferentes graus de diversidade genética entre linhagens, o que permite a escolha de cruzamentos entre progenitores com maior ou menor divergência, conforme os objetivos de determinado programa de melhoramento da soja.

Termos para indexação: Glycine max, distância genética, seleção.

---

* Endereço para correspondência: BIOAGRO, Universidade Federal de Viçosa-UFV, 36570-000 Viçosa, MG.
** Dep. de Fitotecnia da Universidade Federal de Viçosa –UFV- csediyama@ufv.br.
*** Dep. de Biologia Geral da Universidade Federal de Viçosa –UFV.
**** Dep. de Bioquímica e Biologia Molecular da Universidade Federal de Viçosa – UFV - e-mail
Auxílio Financeiro: CNPq.

1. Introdução

O sabor característico da soja, conhecido por "beany flavor", tem sido um dos empecilhos na introdução da soja, como fonte de proteína, na alimentação humana, quando o grão é consumido sem processamento ou quando é processado e utilizado em alimentos industrializados. O "beany flavor" é proveniente da ação catalítica exercida pelas isozimas lipoxigenases (LOX 1, 2 e 3) sobre os ácidos graxos polinsaturados existentes nos grãos de soja (Axelrod, 1974). Os genes responsáveis pela ausência dessas três isozimas são recessivos, ocorrendo estreita ligação entre os genes que codificam LOX 1 e 2, enquanto o gene que codifica LOX 3 segrega independentemente desses dois (Hajika et al., 1991).

A ausência da subunidade protéica A5A4B3 melhora o valor biológico da proteína da soja e confere maior estabilidade à fração protéica, melhorando a qualidade dos produtos industrializados à base de soja (Lanza, 1995). A ausência dessa subunidade é controlada por um gene recessivo, independente dos locos que codificam LOX 1, 2 ou 3.

O programa de melhoramento da soja, em desenvolvimento no Instituto de Biotecnologia Aplicada à Agropecuária - BIOAGRO, da Universidade Federal de Viçosa, visa à obtenção de novos cultivares, com sementes de melhor qualidade e sabor para a alimentação humana (Moreira et al., 1990, Sediyama et al., 1998). Dessa forma, o projeto congrega esforços nas áreas de melhoramento tradicional, bioquímica e genética molecular, dentre outras.

A quantificação da divergência genética visa à identificação de progenitores adequados para o método de melhoramento utilizado. A hibridação de progenitores divergentes geralmente resulta em prole de maior variabilidade, isto é interessante, pois aumenta o número de classes fenotípicas. Em retrocruzamentos, o interesse é a identificação de progenitores menos divergentes, visto que se deseja recuperar as características do progenitor recorrente em menor tempo.

Uma metodologia bastante utilizada para se estimar a divergência genética é a da análise multivariada, em que as distâncias multivariadas são associadas à análise de agrupamento. Dentre as distâncias utilizadas como índice de divergência genética, a distância generalizada de Mahalanobis tem sido empregada nas análises de agrupamento, nos experimentos com repetições.

O objetivo deste trabalho foi determinar a divergência genética entre as linhagens promissoras, previamente selecionadas para ausência de lipoxigenases e da subunidade A5A4B3,

aplicando-se a distância generalizada de Mahalanobis, para auxiliar na decisão da escolha de progenitores destinados à hibridação e obtenção de novas linhagens promissoras.

2. Material genético

O material genético constituiu-se de 91 linhagens promissoras com ausência de lipoxigenases denominadas triplo-nulas (TN), sendo resultantes de cruzamentos entre variedades indicadas para cultivo na região do Brasil Central e progenitores sem lipoxigenases. A caracterização das linhagens, quanto à presença ou ausência de lipoxigenases nas sementes, é realizada por microanálises bioquímicas, não destrutivas, como segue: LOX 1 - teste da oxidação do iodeto (Hammond et al., 1992); LOX 2 - oxidação do sulfato ferroso (Hammond et al., 1992); LOX 3 - co-oxidação do β-caroteno (Kikuchi e Kitamura, 1987); e para confirmação de ausência das três lipoxigenases (LOX 1, 2, 3) - eletroforese em minigéis de poliacrilamida. A detecção da subunidade A5A4B3 é feita utilizando eletroforese em gradiente SDS-PAGE.

3. Delineamento experimental e coleta de dados

O ensaio foi realizado em casa de vegetação, localizado no campus da Universidade Federal de Viçosa, em Viçosa, Estado de Minas Gerais. O experimento constituiu-se de dois blocos completos casualizados, cada um deles contendo 91 tratamentos (linhagens). Cada unidade experimental era constituída por um vaso, com duas plantas de uma mesma linhagem. Foram avaliadas oito características agronômicas, utilizadas para o cálculo da matriz de distância generalizada de Mahalanobis, tomando-se, para os cálculos, as médias das duas plantas. As características agronômicas avaliadas foram as seguintes:

- Número de dias da emergência até o início da floração (estádio R1), sendo denominado número de dias para floração (NPF);

Altura da planta no florescimento (estádio R1), obtida pela medição da distância, em cm, do nível do solo até a extremidade da haste principal (APF);

Número de dias da emergência até a maturação plena (estádio R8), denominado de número de dias para maturação (NDM);

Altura da primeira vagem, obtida pela medição da distância, em cm, do nível do solo até a inserção da primeira vagem da haste principal, na época da colheita (APV);

Número de nós da haste principal na maturação (estádio R8), avaliado a partir do nó correspondente ao par de folhas unifolioladas, até o último nó da haste principal (NNM); e

- Número de vagens com sementes formadas por planta (NVP); número de sementes por planta (NSP) e peso de sementes por planta, ou seja produtividade, em gramas (PSP).

Os termos R1 e R8 referem-se, respectivamente, ao início da floração e à maturação plena, de acordo com Fehr e Caviness (1977).

O modelo estatístico para uma dada característica foi o seguinte:

$$X_{ij} = \mu + t_i + b_j + e_{ij}$$

em que:

X_{ij} = é o valor observado na i-ésima linhagem no j-ésimo bloco;

μ = é a média geral da característica;

t_i = é o efeito do i-ésima linhagem, considerado fixo;

b_j = é o efeito do j-ésimo bloco, considerado aleatório; e

e_{ij} = é o erro experimental associado à observação X_{ij}, em que $e_{ij} \sim N(0, \sigma^2)$.

No Quadro 1, são apresentadas as linhagens promissoras avaliadas com as numerações originais de identificação da linhagem.

3.1. Cálculo da matriz das distâncias generalizadas de Mahalanobis

A distância de Mahalanobis é importante, quando existem repetições dentro das unidades amostrais, e se destaca em relação a outras medidas de dissimilaridade, quando as características avaliadas são correlacionadas, pois leva em consideração a correlação residual entre variáveis.

A matriz de distância generalizada de Mahalanobis (D^2) foi obtida a partir das médias das linhagens, utilizando-se o aplicativo GENES (Cruz, 1997), como segue:

$$D^2_{ii'} = [d_1 \quad d_2 \quad \dots \quad d_n]\, \psi^{-1} \begin{bmatrix} d_1 \\ d_2 \\ \dots \\ d_n \end{bmatrix},$$

$$\psi = \begin{bmatrix} \hat{Var}(X_1) & \hat{Cov}(X_1, X_2) & \dots & \hat{Cov}(X_1, X_n) \\ \hat{Cov}(X_2, X_1) & \hat{Var}(X_2) & \dots & \hat{Cov}(X_2, X_n) \\ \dots & \dots & \ddots & \dots \\ \hat{Cov}(X_n, X_1) & \hat{Cov}(X_n, X_2) & \dots & \hat{Var}(X_n) \end{bmatrix}$$

Quadro 1 - Linhagens promissoras avaliadas

id[1]	Linhagem	id	Linhagem	id	Linhagem
1	Pb2TNG5H 11-01	32	Pb3TNG5H 11-21	63	Pb4TNG5H 5-03
2	Pb2TNG5H 11-2	33	Pb3TNG5H 12-01	64	Pb4TNG5H 5-04
3	Pb2TNG5H 11-03	34	Pb3TNG5H 12-02	65	Pb4TNG5H 5-05
4	Pb2TNG5H 11-04	35	Pb3TNG5H 12-03	66	Pb4TNG5H 5-06
5	Pb2TNG5H 11-05	36	Pb3TNG5H 12-04	67	Pb4TNG5H 5-07
6	Pb2TNG5H 11-06	37	Pb3TNG5H 12-05	68	Pg3TNG5H 3-01
7	Pb3TNG5H 3 1-1	38	Pb3TNG5H 12-06	69	Pg3TNG5H 3-02
8	Pb3TNG5H 3 1-4	39	Pb3TNG5H 13-01	70	Pg3TNG5H 3-03
9	Pb3TNG5H 3 1-5	40	Pb3TNG5H 13-02	71	Pg3TNG5H 4-01
10	Pb3TNG5H 3 2-2	41	Pb3TNG5H 13-03	72	Pg3TNG5H 7-01
11	Pb3TNG5H 3 3-2	42	Pb3TNG5H 13-04	73	Pg3TNG5H 7-02
12	Pb3TNG5H 11-01	43	Pb4TNG5H 2-01	74	Pg3TNG5H 7-03
13	Pb3TNG5H 11-02	44	Pb4TNG5H 2-02	75	Pg3TNG5H 10-01
14	Pb3TNG5H 11-03	45	Pb4TNG5H 2-03	76	Pg3TNG5H 10-02
15	Pb3TNG5H 11-04	46	Pb4TNG5H 2-04	77	Pg3TNG5H 11-02
16	Pb3TNG5H 11-05	47	Pb4TNG5H 2-05	78	Pg3TNG5H 11-03
17	Pb3TNG5H 11-06	48	Pb4TNG5H 2-06	79	1640 TNPb2
18	Pb3TNG5H 11-07	49	Pb4TNH 3-01	80	1642-1 TNPg2
19	Pb3TNG5H 11-08	50	Pb4TNH 3-02	81	1643-5 TNPg2
20	Pb3TNG5H 11-09	51	Pb4TNH 3-03	82	1644-1 TNPg2
21	Pb3TNG5H 11-10	52	Pb4TNH 3-04	83	1649-1 TNPb2
22	Pb3TNG5H 11-11	53	Pb4TNG5H 4-01	84	1649-2 TNPb2
23	Pb3TNG5H 11-12	54	Pb4TNG5H 4-02	85	1649-4 TNPb2
24	Pb3TNG5H 11-13	55	Pb4TNG5H 4-03	86	1650-1 TNPb2
25	Pb3TNG5H 11-14	56	Pb4TNG5H 4-04	87	1657-2 TNPg3
26	Pb3TNG5H 11-15	57	Pb4TNG5H 4-05	88	1660-2 TNPg3
27	Pb3TNG5H 11-16	58	Pb4TNG5H 4-06	89	1641-1 TNPb2
28	Pb3TNG5H 11-17	59	Pb4TNG5H 4-07	90	1649-3 TN Pb2
29	Pb3TNG5H 11-18	60	Pb4TNG5H 4-08	91	16601 TN Pg3
30	Pb3TNG5H 11-19	61	Pb4TNG5H 5-01		
31	Pb3TNG5H 11-20	62	Pb4TNG5H 5-02		

[1] Número de identificação da linhagem.

em que:

$D^2_{ii'}$ = é a distância de Mahalanobis, entre as linhagens i e i';

d_k = é a diferença entre médias de duas linhagens i e i' para uma dada característica k,

isto é, $d_k = \bar{X}_{ik} - \bar{X}_{i'k}$, $k = 1, \ldots, n$; e

ψ = é a matriz de variâncias e covariâncias residuais, de ordem $n \times n$ e simétrica.

Segundo Cruz e Regazzi (1997), para o cálculo de D^2, supõe-se a existência de distribuição normal multivariada n-dimensional e a homogeneidade da matriz de covariância residual das unidades amostrais.

3.2. Análise de agrupamento

A análise de agrupamento refere-se ao processo de arranjar séries de dados em grupos, de forma que os dados de uma série tenham alto grau de homogeneidade (Morrisson, 1967; Johnson e Wichern, 1988). Neste trabalho seguiu o método proposto por Tocher, citado por Rao (1952) e Cruz e Regazzi (1997), que adota o critério de que a média das medidas de dissimilaridade (ex. distância generalizada de Mahalanobis) dentro de cada grupo deve ser menor que as distâncias médias entre quaisquer grupos.

A análise de agrupamento tem o seguinte princípio: seja $\alpha = \{\mu_1, \ldots, \mu_i\}$ um conjunto de características agronômicas e $\beta = \{\gamma_r, \ldots, \gamma_s\}$ o conjunto de elementos para se agrupar. Com base no conjunto α, deve-se determinar uma distribuição dos elementos de β em Γ grupos de tal forma que: se γ_r e $\gamma_s \in \Gamma_j$, então γ_r e γ_s são linhagens convergentes e, se $\gamma_r \in \Gamma_j$ e $\gamma_s \in \Gamma_k$, com $j \neq k$, então γ_r e γ_s são linhagens divergentes. Cada unidade experimental (linhagem) é incluída em um grupo homogêneo se foram semelhantes umas com as outras, senão ficarão em grupos diferentes (Mardia et al., 1979).

4. Resultados e discussões

Os testes não destrutivos para detecção de LOX 1, LOX 2 e LOX 3 nas sementes foram aplicados em amostras formadas por cinco sementes de cada planta, na qual se confirmou a ausência de contaminação do material genético, durante a realização do trabalho. Em razão da grande quantidade de dados coletados, estão apresentados nos quadros apenas os dados de algumas linhagens selecionadas, para ilustrar os dados originais abrangendo o objetivo principal deste trabalho. A pressuposição de normalidade dos erros foi atendida para todas as características estudas pelo teste de normalidade de Shapiro-wilk (1965), por meio do software SAS procedimento Proc Univariate, utilizando a opção Normal.

Observa-se, no Quadro 2, que houve grande amplitude de variação nas características agronômicas, sendo as maiores médias de peso de sementes por planta (PSP) obtidas com as

linhagens Pb3TNG5H 11-06, Pb4TN 3-02, Pb4TNG5H 4-05 e Pb4TNG5H 5-05. Observa-se, ainda, a presença de linhagens de diferentes grupos de maturação entre as mais produtivas. Esse fato é interessante pois tem-se a possibilidade de selecionar linhagens mais precoces ou mais tardias, possibilitando melhor gerenciamento da produção da propriedade rural, com relação à flexibilidade de recomendação de variedades (Sediyama et al., 1985).

Com base na matriz das distâncias generalizadas de Mahalanobis, apresentadas no Quadro 3, observa-se que as linhagens Pg3TNG5H 10-02 e 1640 TNPb2 foram as mais convergentes entre si e as linhagens Pb3TNG5H 11-03 e Pg3TNG5H 10-01 foram as mais divergentes. Hibridação entre progenitores mais divergentes possibilitam o desenvolvimento de linhagens mais produtivas que os seus progenitores devido a sua ampla diversidade genética (Naoe, 2001), enquanto progenitores mais convergentes são utilizados para retrocruzamento, onde se visa a recuperar rapidamente uma ou mais características agronômicas de interesse econômico (Moreira et al., 1996).

Quadro 2 - Médias das características agronômicas de soja de algumas linhagens promissoras selecionadas

id[1]	Linhagem	NPF[2]	APF	NDM	APV	NNM	NVP	NSP	PSP
14	Pb3TNG5H 11-03	61	114	118	29	13	69	145	30
17	Pb3TNG5H 11-06	60	125	130	28	14	93	217	52
50	Pb4TN 3-02	63	111	132	23	16	100	195	47
57	Pb4TNG5H 4-05	57	111	117	26	14	78	248	43
65	Pb4TNG5H 5-05	59	107	117	17	13	78	174	42
75	Pg3TNG5H 10-01	43	80	110	28	9	38	79	19
76	Pg3TNG5H 10-02	62	130	133	26	14	70	152	30
79	1640 TNPb2	62	124	137	24	13	55	120	23

[1] Número de identificação da linhagem.
[2] Siglas descritas no item 3.

Quadro 3 - Parte da matriz de distância generalizada de Mahalanobis das linhagens promissoras

Linhagem	17[1]	50	57	65	75	76	79
14. Pb3TNG5H 11-03	28,607	40,145	24,472	33,274	82,453	13,995	16,879
17. Pb3TNG5H 11-06	-	13,773	13,570	10,792	35,560	12,427	13,729
50. Pb4TN 3-02		-	19,471	17,361	55,774	14,137	15,715
57. Pb4TNG5H 4-05			-	18,319	47,202	21,072	23,573
65. Pb4TNG5H 5-05				-	41,272	21,285	22,051
75. Pg3TNG5H 10-01					-	52,067	52,437
76. Pg3TNG5H 10-02						-	1,010
79. 1640 TNPb2							-

[1] Número de identificação da linhagem.

No Quadro 4, são apresentados os resultados do agrupamento pelo método de otimização de Tocher, com base na matriz de distância generalizada de Mahalanobis, das 91 linhagens promissoras de soja com ausência de lipoxigenases 1, 2 e 3 e da subunidade A5A4B3. A análise de agrupamento discriminou dezessete grupos distintos, sendo os grupos I, II, III, IV, V, VI, VII, VIII e IX constituídos por 26, 16, 6, 5, 6, 4, 3, 2, 2 linhagens, respectivamente, e os demais grupos constituídos por apenas uma linhagem.

As hibridações indicadas para exploração de variabilidade genética é a intergrupo e a de recuperação de características a intragrupos. Um exemplo de hibridação é descrito por 'Pb3TNG5H 11-06 x Pb4TN 3-02' uma vez que são linhagens produtivas e foram alocadas em grupos diferentes.

Quadro 4 - Resultado do agrupamento pelo método de otimização de Tocher das 91 linhagens promissoras de soja com ausência de lipoxigenases 1, 2 e 3 e da subunidade A5A4B3

Grupos distintos	Número de identificação da linhagem
I	2, 3, 5, 8, 10, 11, 17, 19, 20, 32, 34, 35, 36, 37, 44, 45, 46, 47, 48, 49, 52, 54, 55, 59, 60, 61, 62, 64, 70, 71, 74, 76, 79, 82, 83, 85, 89, 90 e 91
II	7, 18, 28, 29, 33, 38, 39, 40, 41, 42, 43, 67, 72, 73, 75 e 77
III	9, 51, 57, 65, 66 e 78
IV	6, 12, 15, 31 e 68
V	4, 13, 84, 86, 87 e 88
VI	50, 53, 58 e 63
VII	25, 26 e 27
VIII	16 e 30
IX	14 e 22
Linhagens distintas	1, 21, 23, 24, 56, 69, 80 e 81

No caso das oito linhagens que formaram oito grupos distintos, podem constituir-se em materiais de diferenciados potenciais, quanto ao melhoramento da soja, sugerindo um cuidadoso estudo de cada linhagem, para melhor indicação de hibridação, pois algumas delas podem significar material genético com características interessantes ao produtor rural.

5. Conclusões

Com base nos resultados, conclui-se que:

a) O material genético avaliado não apresentou segregação gênica ou contaminação, quanto à ausência de lipoxigenases ou da subunidade protéica A5A4B3;

b) A análise de agrupamento identificou dezessete grupos distintos no material genético analisado; e

c) Há diferentes graus de diversidade genética entre linhagens, o que permite a escolha de cruzamentos entre progenitores com maior ou menor divergência, para o desenvolvimento de novas linhagens superiores com sementes ausentes de lipoxigenases e da subunidade A5A4B3.

Referência bibliográfica

AXELROD, B. (1974). Lipoxygenases. Advanced Chemical Series, v.2, p.342-348.

CRUZ, C.D. (1997). Programa GENES. Viçosa - MG: Editora UFV, 442p.

CRUZ, C.D. & REGAZZI, A. J. (1997). Modelos biométricos aplicados ao melhoramento genético. Viçosa - MG: Editora UFV, 390p.

FEHR, W.R.. & CAVINESS, C.E. (1977). Stages of soybean development. Ames, Iowa: Cooperative Extension Service; Iowa State University. 11p. (Special Report, 80).

HAJIKA, M.; IGITA, K. & KITAMURA, K. (1991). A line lacking all the seed lipoxygenase isozymes in soybean (Glycine max (L.) Merrill) induced by gamma-ray irradiation. Japanese Journal of Breeding, v.41, p.507-509.

HAMMOND, E.G.; DUVICK, D.N.; FEHR, W.R.; HILDEBRAND, D.F.; LACEFIELD, E.C. & PFEIFFER, T.W. (1992). Rapid screening techniques for lipoxygenases in soybean seeds. Crop Science, v.32, p.820-821.

JOHNSON, R. & WICHERN, D.W. (1988). Applied multivariate statistical analysis. 2 ed. New Jersey: Prentice-Hall, 607p.

KIKUCHI, A. & KITAMURA, K. (1987). Simple and rapid carotene bleaching tests for the detection of lipoxygenase isozymes in soybeans seeds. Japanese Journal of Breeding. v.37, p.10-16.

LANZA, M.A. (1995). Marcadores moleculares RAPD na introgressão de genes para ausência de lipoxigenases e da proteína A5A4B3 em soja (Glycine max (L.) Merrill). Viçosa - MG: UFV. 57p. (Tese de Mestrado).

MARDIA, K.V.; KENT, J.T.; BIBBY, J.M. Multivariate analysis. London: Academic Press, 1979. 518p.

MOREIRA, M. A.; BARROS, E. G.; SEDIYAMA, C. S.; SEDIYAMA, T. (1996). Breeding soybean for high quality seeds assisted by molecular markers. In: PLANT GENOME IV CONFERENCE. San Diego, CA, USA.

MOREIRA, M.A., REZENDE, S.J., SEDIYAMA, C.S. & GOMES, J.C. (1990). Obtenção de cultivares de soja de sabor agradável e com sementes de alta qualidade fisiológica. In: TORRES, A.C., CALDAS, L.S. (Eds.). (1990). Técnica e aplicações da cultura de tecidos de plantas. Brasília - DF: ABCTP. p. 417-426.

MORRISSON, D.F. (1967). Multivariate statistical methods. New York: McGraw-Hill, 415p.

NAOE, L.K.; SEDIYAMA, C.S.; MIRANDA, G.V.; CRUZ, C.D. & MOREIRA, M.A. (2001). Diversidade entre progenitores e variabilidade das populações segregantes de soja. Revista Ceres. v.48, p.223-237.

RAO, R.C. (1952). Advanced statistical methods in biometric research. New York: John Wiley and Sons, 390p.

SEDIYAMA, C.S.; QUEIROZ, L.R.; MOREIRA, M.A. & REZENDE, S.T. (1998). Aldehyde production and physiological quality of soybean seeds lacking lipoxygenase isozymes. In: WORLD SOYBEAN RESEARCH CONFERENCE, V, Chiang Mai, Thailand. Proceedings... (Supplement), Kasetsart University Press, Bangkok. p.441-446.

SEDIYAMA, T.; PEREIRA, M.G. ; SEDIYAMA, C.S. & GOMES, J.L.L. (1985). Cultura da soja – I parte. Viçosa - MG: Universidade Federal de Viçosa. 96p.

SHAPIRO, S.S. & WILK, M.B. (1965). An analysis of variance test for normality. Biometrika, v.52, p.591-611.

Abstract

This work was carried out with potentially interesting soybean inbred lines derived from a breeding program aiming at the production of new soybean cultivars with grains suited to human consumption. The lines were previously selected for all the three lipoxygenases and the A5A4B3 protein subunit absence. Eight agronomic characters were used to indicate distinct groups for crossing, based on the Tocher´s clustering method, using the genetic divergence between inbred lines estimated by the multivariate technique of Mahalanobis generalized distance. According to the obtained results, it was concluded that there are different degrees of genetic diversity among the inbred lines, which allow the selection of progenitors with greater or smaller divergence to meet the needs of a given program of soybean breeding.

Index terms: Glycine max, genetic distance, selection

Estimação consistente de matrizes de covariâncias sob heteroscedasticidade de forma desconhecida: um estudo de simulação

Raydonal Ospina*
Bartolomeu Zamprogno*
Patrícia Leone Espinheira*

Resumo

Neste artigo, nós examinamos o comportamento de alguns conhecidos estimadores consistentes da matriz de covariância do estimador de mínimos quadrados ordinários sob heteroscedasticidade e propomos alguns novos estimadores. Toda nossa análise é baseada em simulações de Monte Carlo, onde verificamos o comportamento do viés relativo total e do erro quadrático médio total destes estimadores. Além disso, observamos o desempenho dos testes quasi-*t* baseados nos diferentes estimadores, em dois cenários diferenciados segundo a ocorrência de observações de alta alavancagem.

Palavras chaves: Heteroscedasticidade, teste quasi-*t*, pontos de alta alavancagem.

* Departamento de Estatística, CCEN, Universidade Federal de Pernambuco, Cidade Universitária, Recife/PE, 50740-540, Brasil.

1. Introdução

O modelo de regressão linear é comumente usado por econometristas e estatísticos aplicados. Este modelo, junto a outras generalizações, constitui a base da maioria dos trabalhos empíricos em economia e áreas afins. A técnica clássica de regressão linear exige a suposição de que os erros têm variância constante, que em geral não é muito plausível. Por exemplo, dados de corte transversal tipicamente apresentam alguma forma de heteroscedasticidade, isto é, as variâncias dos erros não são iguais para todas as observações. Quando a forma e a magnitude da heteroscedasticidade são conhecidas, usando-se ponderações adequadas é possível corrigi-la, e, assim, é possível usar mínimos quadrados generalizados. No caso em que a forma da heteroscedasticidade envolve um pequeno número de parâmetros desconhecidos, podemos realizar estimações do modelo de regressão através do estimador de mínimos quadrados generalizados viável, que pressupõe uma estrutura funcional da forma da heteroscedasticidade. Em muitos casos, entretanto, dita estrutura funcional é desconhecida, o que torna impraticável o uso de ponderações aproximadas. Mesmo quando a suposição de variância constante é violada, o estimador de mínimos quadrados ordinários (EMQO) permanece não-viesado e consistente, tornando-se, contudo, ineficiente; o estimador usual da matriz de covariância do EMQO torna-se viesado e deixa de ser consistente. O interesse principal de muitos autores tem recaído sobre a obtenção de estimadores consistentes da matriz de covariância do EMQO com o fim de melhorar a qualidade e precisão das inferências sobre os parâmetros da regressão. O estimador consistente da matriz de covariância sob heteroscedasticidade (HCCME) mais comumente utilizado foi proposto por Halbert White em 1980, o qual é uma reconstrução do estimador proposto por Eicker (1963) e que é denominado HC0[1]. O artigo onde HC0 foi proposto originalmente tem aproximadamente 2 600 (duas mil e seiscentas) citações de acordo com o Instituto de Informações Científicas; indicando que o artigo de White ocasionou um profundo impacto na literatura de estimação de modelos de regressão. Entretanto, estudos de Monte Carlo evidenciam que o estimador de White pode ser consideravelmente viesado em amostras finitas, tendendo a subestimar as variâncias verdadeiras, tornando assim o teste quasi-t associado muito liberal; ver MacKinnon e White (1985), Cribari-Neto e Zarkos (1999, 2001) e Cribari-Neto, Ferrari e Cordeiro (2000).MacKinnon e White (1985) consideraram HCCMEs alternativos e acharam que o estimador da matriz de covariância obtido por Jackknife tem performance superior à de outros

[1] O estimador de White está implementado em alguns pacotes estatísticos, tais como SHAZAM.

estimadores em amostras de tamanho finito, incluindo o estimador de White. Davidson e MacKinnon (1993) argumentam que o estimador de Jackknife é bem aproximado por uma variante do estimador de White conhecida como HC3. Long e Ervin (2000) realizaram extensivas simulações e os resultados obtidos mostram que o estimador HC3 tem desempenho superior ao de outras variantes do estimador de White (1980) e também ao estimador originalmente proposto por White. Baseando-se em experimentos de Monte Carlo, MacKinnon e White (1985) recomendam o uso do estimador HC3 em pequenas amostras[2]. A partir de outras simulações, Davidson e MacKinnon (1993) recomendam em geral o uso dos estimadores HC2 e HC3 em detrimento do estimador HC0. Resultados em Cribari-Neto e Zarkos (2001) mostram que a presença de pontos de alta alavancagem na matriz de regressores é mais decisiva para o comportamento de HCCMEs do que o grau de heteroscedasticidade em si, quando o tamanho da amostra é finito. Os testes quasi-t associados tendem a ser muito liberais quando os dados incluem observações de alta alavancagem, tornando, assim, a inferência imprecisa. Cribari-Neto (2004) propõe um HCCME que leva em consideração o impacto dos pontos de alta alavancagem sobre o comportamento, em amostras finitas, do estimador da matriz de covariância e denomina este novo estimador de HC4. Este novo estimador tem um bom desempenho em amostras de tamanho finito, quando é usado na construção do teste quasi-t e quando os dados possuem observações de alta alavancagem.

Neste trabalho, apresentamos, na seção 2, os conceitos básicos em estudo e as definições dos estimadores consistentes da matriz de covariância sob heteroscedasticidade de forma desconhecida. Algumas propostas novas de estimadores são apresentadas na seção 3 e um algoritmo para determinar o grau de heteroscedasticide. Nas seções 4 e 5 avaliamos o comportamento destes estimadores, onde verificamos que os estimadores HC8 e HC9 apresentam, em geral, os melhores desempenhos em relação aos resultados dos testes quasi-t, quando existem pontos de alavanca médios e fortes.

2. O modelo e estimadores

Seja o modelo de regressão linear clássico da forma

$$y = X\beta + \varepsilon$$

[2] Long e Ervin (2000) pesquisaram 12 pacotes estatísticos e mostram que o mais comumente implementado estimador consistente da matriz de covariância sob heteroscedasticidade é o estimador de White (HC0), enquanto o estimador HC3 somente está implementado no STATA e no TSP.

onde y é um vetor de n observações da variável dependente, X é uma matriz nxp , de posto p, de observações das variáveis explicativas, β é um vetor de p parâmetros fixos e desconhecidos e ε um vetor de erros aleatórios não observáveis. Consideremos as seguintes suposições:

[S1] O modelo proposto é o verdadeiro;

[S2] $E(\varepsilon_i) = 0, i = 1,...,n;$

[S3] $Var(\varepsilon_i) = E(\varepsilon^2_i) = \sigma^2, i = 1,..., n;$

[S4] $Cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall i = j \ \ i,j = 1,..., n;$

[S5] $\left(\frac{X'X}{n}\right) \to Q,$ quando $n \uparrow \infty$ onde Q é uma matriz finita e positiva definida.

As suposições S2 e S4 indicam que os erros além de ter média zero, são não correlacionados e a suposição S3 implica que o modelo é homoscedástico, ou seja, as variâncias dos erros são iguais para todas as observações. Assim, temos que, $Cov(\varepsilon) = \Omega = \sigma^2 I_n$, onde I_n é a matriz identidade de ordem n. A suposição S5 é necessária para garantir as propriedades assintóticas dos estimadores do modelo. O principal interesse da modelagem de regressão é estimar o vetor de parâmetros e fazer inferências sobre os mesmos. O método de mínimos quadrados ordinários (MQO) fornece o seguinte estimador:

$$b = (X'X)^{-1}X'y.$$

Duas outras quantidades de interesse podem ser encontradas a partir de b; o vetor de valores preditos e o vetor de resíduos de MQO, os quais são definidos respectivamente, como

$$\hat{y} = Xb = X(X'X)^{-1}X'y = Hy,$$
$$\hat{e} = y - \hat{y} = y - Hy = My,$$

onde $H = X(X'X)^{-1}X'$ e $M = I_n - X(X'X)^{-1}X'$ são matrizes simétricas e idempotentes.

Podemos mostrar que b é um estimador não-viesado e consistente para β, isto é quando n torna-se grande, b aproxima-se do parâmetro com probabilidade crescente. Ressaltamos que a hipótese de homoscedasticidade não é necessária para garantir ditas propriedades. No entanto, existe uma propriedade amostral de b que está diretamente ligada à estrutura das variâncias dos

erros: a eficiência. O teorema de Gauss-Markov (Judge et al., 1988, p. 202) atesta que a variância de b dada por:

$$Cov(b) = \sigma^2 (X'X)^{-1}$$

é mínima na classe de estimadores lineares e não-viesados para β, implicando a eficiência de b. Desta forma, como conseqüência do teorema de Gauss-Markov temos que o EMQO é dito ser BLUE (*Best Linear Unbiased Estimator*) para o vetor de parâmetros quando $Cov(\varepsilon) = \sigma^2 I_n$. Assim, em um modelo de regressão linear homoscedástico em que os erros têm média zero, o EMQO de β é não-viesado, consistente e eficiente.

Quando as variâncias dos erros são não constantes, o modelo é dito ser heteroscedástico e a propriedade de eficiência é a única afetada. Desta forma, a matriz de covariância de ε toma a forma

$$Cov(\varepsilon) = \Omega = diag(\sigma_1^2, \ldots, \sigma_n^2).$$

Se Ω for conhecido, podemos transformar o modelo original em um modelo homoscedástico e estimar seus parâmetros através de mínimos quadrados generalizados (EMQG). No entanto, na prática a forma da heteroscedasticidade é desconhecida e o EMQG não é adequado, dependendo da estrutura funcional da forma de heteroscedasticidade assumida. Assim, uma alternativa é tentar estimar Ω e construir, por exemplo, o estimador de mínimos quadrados generalizados viável (EMQGV), que precisa assumir uma forma funcional para a variância dos erros, o que em geral pode tornar imprecisas as inferências resultantes sobre os parâmetros do modelo. Assim, o fato da variância dos erros não ser constante afeta as inferências sobre o vetor de parâmetros. Uma vez que o EMQO de β continua não-viciado e consistente, o problema a ser resolvido é a estimação da matriz de covariância de b, dada por:

$$\Psi = Cov(b) = (X'X)^{-1} X' \Omega X' (X'X)^{-1},$$

o que aparentemente requereria estimar Ω. Halbert White em 1980 propõe um estimador para Cov(b) da forma

$$\Psi = Cov(b) = (X'X)^{-1} X' \hat{\Omega} X' (X'X)^{-1},$$

onde $\hat{\Omega} = diag(\ddot{e}_1^2, \ldots, \ddot{e}_n^2)$. Este estimador se apresenta consistente tanto sob homoscedasticidade quanto sob heteroscedasticidade de forma desconhecida. MacKinnon e White (1985) mostram que o estimador proposto por White pode ser muito viesado em amostras finitas, já que os resíduos tendem a subestimar os erros. O fato dos resíduos de MQO serem pequenos quando comparados aos erros está diretamente ligado aos pontos de alavanca. Estes pontos forçam a reta de regressão a se aproximar deles diminuindo o seu resíduo, induzindo assim que a respectiva variância estimada seja menor do que deveria.

Cribari-Neto e Zarkos (2001) mostraram que a presença de observações com alta alavancagem pode ser mais decisiva para o comportamento do estimador HC0 em amostras finitas que o grau de heteroscedasticidade em si. Uma metodologia para detectar observações com alta alavancagem é observando os elementos diagonais h_i da matriz H. Segundo Davidson e MacKinnon (1993, §1.6), quando h_i é relativamente grande, a observação relacionada a este h_i tem um efeito relativamente grande sobre b, a não ser que seu respectivo resíduo seja próximo de zero. Desta forma, a quantidade h_i pode ser usada como medida de alavancagem ou de efeito potencial desta observação sobre b. Se existirem observações para as quais h_i é maior que $2p/n$ ou $3p/n$, dizemos que estas têm alta alavancagem. O desempenho de estimador de White é mais afetado em relação ao teste quasi-t. No sentido de que ele tende a subestimar as variâncias verdadeiras e os testes associados tendem a ser muito liberais, ou seja, rejeitam a hipótese nula quando de fato ela e verdadeira mais freqüentemente do que deveriam. Suponhamos que o objetivo é testar alguma suposição a respeito de um parâmetro específico do modelo, por exemplo, $H0$: $\beta k = m$ contra $H1$: $\beta k \neq m$, onde βk é o k-ésima coeficiente de uma dada regressão e m é uma dada constante. A estatística de teste é dada por $\tau = (\beta_k - m) / \sqrt{\hat{var}(b_k)}$, onde, $\hat{var}(b_k)$ é o elemento (k,k) de um HCCME Independente da distribuição dos erros, e mesmo sob heteroscedasticidade, temos o seguinte resultado:

$$\tau \xrightarrow{d} N(0,1) \text{ quando } n \to \infty .$$

Assim, podemos usar os valores críticos aproximados de uma distribuição limite normal padrão para realizar o teste de hipótese, o teste e denominado de "teste quais-t".

Uma das primeiras tentativas de melhorar o desempenho do estimador HC0 foi proposta por Hinkley (1977); sua idéia consistiu em corrigir o viés negativo ponderando o estimador HC0 por $n/(n-p)$, i.e., o estimador para Ω é dado por

$$\hat{\Omega} = \frac{n}{n-p} diag\left(\tilde{e}_1^2, \ldots, \tilde{e}_n^2\right).$$

Este estimador é conhecido como HC1. Outras variações foram propostas com o fim de tentar diminuir o viés do estimador de White e melhorar o seu desempenho em relação aos testes "quasi-t". Baseado em Horn, Horn e Duncan (1975), MacKinnon e White (1985) sugerem um estimador conhecido como HC2, o qual estima Ω da seguinte forma:

$$\hat{\Omega} = \frac{n}{n-p} diag\left(\tilde{e}_1^2, \ldots, \tilde{e}_n^2\right).$$

Este estimador é conhecido como HC1. Outras variações foram propostas com o fim de tentar diminuir o viés do estimador de White e melhorar o seu desempenho em relação aos testes "quasi-t". Baseado em Horn, Horn e Duncan (1975), MacKinnon e White (1985) sugerem um estimador conhecido como HC2, o qual estima Ω da seguinte forma:

$$\hat{\Omega} = diag\left(\frac{\tilde{e}_1^2}{(1-h_1)}, \ldots, \frac{\tilde{e}_n^2}{(1-h_n)}\right).$$

Este estimador, sob homoscedasticidade se apresenta não-viesado. MacKinnon e White (1985) propõem o estimador HC3, que estima Ω através da expressão:

$$\hat{\Omega} = diag\left(\frac{\tilde{e}_1^2}{(1-h_1)^2}, \ldots, \frac{\tilde{e}_n^2}{(1-h_n)^2}\right).$$

O estimador HC3 apresenta propriedades muito semelhantes ao estimador Jackknife e oferece desempenho melhorado, tanto em termos do viés quanto aos resultados dos testes "quasi-t". Long e Ervin (2000) sugerem o uso deste estimador para amostras de tamanho menor que 250.

Cribari-Neto (2003) mostra que o estimador HC3 continua sendo muito na presença de observações com grau de alavancagem muito alto. Assim, uma sugestão consiste em ponderar os resíduos, associando de maneira mais eficiente o grau de alavancagem.

A partir desta percepção, Cribari-Neto (2004) propõe o estimador denominado HC4, que utiliza

$$\hat{\Omega} = diag\left(\frac{\hat{e}_1^2}{(1-h_1)^{\delta_1}}, \ldots, \frac{\hat{e}_n^2}{(1-h_1)^{\delta_n}} \right),$$

onde $\delta_i = min\{4, hi / \bar{h}\}$, i = 1,...,n, e $\bar{h} = p/n$. Resultados de simulação revelam que o estimador HC4 apresenta o melhor desempenho em relação ao teste "quasi-*t*", no entanto tende a superestimar as variâncias verdadeiras.

3. Novos estimadores

Evidência empírica mostra que os HCCMEs descritos anteriormente têm um desempenho afetado devido à presença de pontos de alta alavancagem. Isto, em geral, acontece porque as variâncias verdadeiras dos erros são subestimadas. Considerando que mesmo sob homoscedasticidade $E(\hat{e}_i^2) = (1-h_i)\sigma^2 \neq \sigma^2$, uma idéia para que os HCCMEs se apresentem mais eficientes é encontrar uma forma funcional adequada que envolva os resíduos e as quantidades h_i que permita reduzir o viés e otimize os resultados dos testes "quasi-*t*" associados. Neste contexto, nos propomos algumas versões de HCCMEs cuja forma funcional geral é:

$$\hat{\Omega}_{HC_J} = diag\left(\frac{\hat{e}_1^2}{(1-h_1)^{\delta_{J_1}}}, \ldots, \frac{\hat{e}_n^2}{(1-h_n)^{\delta_{J_n}}} \right).$$

Para, $J = 0,...,4$ são estimadores H0 ate HC4 conhecidos, se $J = 5,...,9$ e $i = 1,...,n$ consideramos versões do estimador HC4 da forma

$$HC5: \ \delta_{5_i} = min\{2.5, h_i / \bar{h}\},$$
$$HC6: \ \delta_{6_i} = min\{3.0, h_i / \bar{h}\},$$
$$HC6: \ \delta_{7_i} = min\{3.5, h_i / \bar{h}\},$$
$$HC8: \ \delta_{8_i} = min\{4.5, h_i / \bar{h}\},$$
$$HC9: \ \delta_{9_i} = min\{5.0, h_i / \bar{h}\}.$$

Para $J=10$, consideramos um estimador que se baseia na esperança do quadrado dos resíduos quando as variâncias dos erros não são constantes. Neste caso, $E(\hat{e}_i^2) = (1-h_i(\psi_i - 2))\sigma^2$ (Belsley, 2002), onde

$$\psi_i = \frac{\sum_{i=1}^{n} h_i^2 \sigma_i^2}{h_i \sigma_i^2},$$

Assim, o estimador de Ω é da forma:

$$\hat{\Omega}_{HC_J} = diag\left(\frac{\hat{e}_1^2}{(1-h_1(\psi_1-2))}, \ldots, \frac{\hat{e}_n^2}{(1-h_n(\psi_n-2))}\right) \text{ e } \psi_i = \frac{\sum_{i=1}^{n} h_i^2 \hat{e}_i^2}{h_i \hat{e}_i^2}.$$

As propriedades amostrais do estimador HC10 não foram avaliadas neste trabalho, logo não garantimos que este seja um HCCME, no entanto, achamos de interesse avaliar seu desempenho. Propomos também um estimador que denominamos HC11 considerando $\delta_{11} = \min_i \left\{ 1/(\bar{h} - h_i) \right\}$ fixo para todas as observações. Já, para $J = 12,\ldots,18$ e $i = 1,\ldots,n$ temos

$$\begin{aligned}
&\text{HC12}: \delta_{12_i} = 1.5,\\
&\text{HC13}: \delta_{13_i} = 2.5,\\
&\text{HC14}: \delta_{14_i} = 3.0,\\
&\text{HC15}: \delta_{15_i} = 3.5,\\
&\text{HC16}: \delta_{16_i} = 4.0,\\
&\text{HC17}: \delta_{17_i} = 4.5,\\
&\text{HC18}: \delta_{18_i} = 5.0.
\end{aligned}$$

Dado $\Omega = diag(\sigma_1^2,\ldots, \sigma_n^2)$, se definimos a matriz P de dimensão pxp da forma $P = (X'X)^{-1}$ e

$$D_{HC_J} = diag\left(\frac{1}{(1-h_1)^{\delta_{J1}}}, \ldots, \frac{1}{(1-h_1)^{\delta_{Jn}}}\right),$$

então, Cribari-Neto e Galvão (2003) mostram, que, em geral, o viés do estimador $\hat{\Psi}$ para qualquer escolha de D_{HC_J} é da forma :

$$B(\hat{\Psi}) = P[D_{HC_J}\{H\Omega H - 2H\Omega\} + D_{HC_J}]P',$$

Desta forma, este tipo de estimadores podem ser corrigidos por viés para melhorar seu desempenho na presença de heteroscedasticidade. Nosso trabalho concentra-se na avaliação numérica dos estimadores propostos sem correção por viés.

Na prática, detectada a presença de heteroscedasticidade, por exemplo, através de um teste de Goldfel-Quandt ou Breusch-Pagan (Judge et al., 1988, p. 372) se esta interessado em determinar o grau de heteroscedasticidade λ, o qual é definido por alguns autores, como $\lambda = max(\sigma_t^2)/min(\sigma_t^2)$, $t = 1,...,n$. Com esta idéia em mente nos propomos o seguinte algoritmo para medir o grau de heteroscedasticidade, baseados em que $E(\hat{e}_i^2) = (1 - h_i)\sigma^2$, e $\frac{E(\hat{e}_i^2)}{(1-h_i)} = \sigma^2$.

Seja $b^* = b_{(t)}$ o estimador de β obtido ao calcular o EMQO sim o t-ésimo par (y_t,x_t) da amostra, onde x_t é a t-ésima linha de X , e y^* a variável resposta sim o t-esimo elemento.

Algoritmo λ^*

Para $t = 1,...,n$.

Calcule b^*, $e^* = y^* - X^* b^*$, e $s^* = e^{*\prime} e^*/(n-p)$.

Defina $\sigma_t^{*2} = (1-h_t)\, s^*$.

Calcule $\lambda^* = max(\sigma_t^{*2})/min(\sigma_t^{*2})$.

Retorne λ^*.

Assim, λ^* é uma quantidade que mede o grau de heteroscedasticidade presente nos dados do modelo de regressão. Em nosso trabalho não estudaremos o comportamento de λ^* já que para a análise dos estimadores nós controlamos o grau de heteroscedasticidade.

4. Avaliação numérica

Nesta seção, nós avaliamos o comportamento dos estimadores propostos e os compararmos com os estimadores HCCMEs já conhecidos. Os resultados numéricos nesta secção referem-se ao modelo

$$y_i = \beta_1 + \beta_2 x_i + \sigma_i \varepsilon_i, \quad i = 1,...n.$$

A geração dos dados considera $\beta 1 = \beta 2 = 1$ e os erros, $\varepsilon i's$, independentes e identicamente distribuídos de acordo com uma distribuição N(0,1). Os tamanhos amostrais usados são n = 50, 100, 150, 200. São geradas 50 observações para x de uma distribuição exponencial, com o objetivo de avaliar o impacto dos pontos de alavanca e, para comparar o

desempenho dos estimadores quando não ocorrem tais pontos, são geradas observações de uma distribuição uniforme. Os tamanhos amostrais são obtidos, replicando os 50 valores iniciais de xi, desta forma o grau de heteroscedasticidade, que pode ser medido por $\lambda = max(\sigma t2)/min(\sigma t2)$, mantém-se constante à medida que o tamanho da amostra cresce. O mecanismo gerador da heteroscedasticidade é da forma

$$\sigma_i^2 = exp(\gamma + \gamma x_i^2).$$

Desta forma, para $\gamma = 0$, temos $\lambda = 1$, o que representa homoscedasticidade, já para $\lambda > 1$ temos heteroscedasticidade e os maiores valores de γ, tanto na exponencial quanto na uniforme, conduzem a um grau de heteroscedastididade no máximo igual a 100. Todos os experimentos de Monte Carlo são baseados em 10 000 (dez mil) réplicas e são implementados na linguagem de programação Ox (Doornik, 2001).

As Tabelas de 1 até 3 sumarizam os resultados desta investigação considerando a distribuição uniforme, isto é, considerando o modelo sim pontos de alta alavancagem. Na Tabela 1, estão apresentados os vieses relativos totais dos seguintes estimadores: de mínimos quadrados ordinários (MQO), que tem a forma $\hat{\sigma}^2 (X'X)^{-1}$, onde $\hat{\sigma}^2 = \hat{e}'\hat{e}/(n-p)$, de White (HC0), de Hinkley (HC1), o estimador adaptado por MacKinnon e White (HC2), o estimador jackknife aproximado (HC3), o estimador proposto por Cribari-Neto (HC4), além das novas propostas, (HCJ), $J = 5,...18$, apresentadas na seção anterior. Define-se viés relativo total como a soma dos valores absolutos dos vieses relativos para as variâncias estimadas de $\hat{\beta}_1$ e $\hat{\beta}_2$.

Assim, para cada estimador foi estimado

$$\frac{|E\{\hat{var}(\hat{\beta}_1) - var(\hat{\beta}_1)\}|}{var(\hat{\beta}_1)} + \frac{|E\{\hat{var}(\hat{\beta}_2) - var(\hat{\beta}_{21})\}|}{var(\hat{\beta}_2)},$$

onde "$\hat{var}$" denota o estimador da variância de interesse.

As Tabelas de 1até 3 também apresentam os resultados variando o grau de heteroscedasticidade e o tamanho da amostra, contudo, sem presença de pontos de alavanca, devido à natureza da geração dos dados. De acordo com os dados dispostos na Tabela 1, como era de se esperar, o estimador MQO é não-viesado quando as variâncias dos erros são constantes e consideravelmente viesado sob heteroscedasticidade. É possível observar também que, para todos os tamanhos amostrais e os diferentes graus de heteroscedasticidade, o estimador HC2

apresenta o menor viés relativo total entre todos os HCCMEs avaliados. O segundo melhor desempenho, dentre as novas propostas quanto a este critério, foi apresentado, em geral, pelo estimador HC12, seguido pelo estimador HC10. Se fixamos o tamanho de amostra em 50 e $\gamma = 4$, o qual resulta em $\lambda = 50.44$, encontramos que o viés do estimador HC2 é de 2.64%, do estimador HC12, 4.53% e do estimador HC10, 7.91%. Os estimadores HC4 a HC9 apresentaram resultados idênticos, já que sem a presença de pontos de alavanca, os h_i's são em média menores que p/n, então $\delta_J = \min\{k, h_i/\bar{h}\} = h_i/\bar{h}$ *para* $k > 1$ *e* $J = 4,...,9$. Quando analisamos os resultados para $n = 150$ e $\gamma = 4$, o viés relativo total de HC9 e igual a 2.072%.

Quanto aos estimadores que utilizaram um expoente fixo para a quantidade $(1- h_i)$, à medida que aumenta o valor do expoente δ_J o desempenho é bastante prejudicado, pois estes estimadores tendem a subestimar a variância do erro. Por exemplo, para $n = 200$ e $\gamma = 4$, os vieses relativos totais dos estimadores HC13, HC14, HC15, HC16, HC17 e HC18 são, respectivamente, 4.56%, 6.01%, 7.47%, 8.94%, 10.42% e 11.92%. Os erros quadráticos médios totais dos diferentes estimadores das variâncias são apresentados na Tabela 2. Quanto ao critério do erro quadrático médio (EQM) total, o estimador MQO se comporta melhor que o estimador HC0, para tamanho de amostras inferiores a 100, caso contrário o estimador de White é o que apresenta o melhor desempenho entre todos os HCCMEs. Entre as novas propostas encontramos que os menores erros quadráticos médios se referem ao estimador HC10. Novamente, os estimadores HC4 até HC9 se comportam de forma idêntica. Quanto aos estimadores que utilizam expoentes fixos para a quantidade $(1- h_i)$, à medida que o expoente cresce o desempenho piora.

A Tabela 3 contém as taxas de rejeição da hipótese nula (em percentuais), para um nível nominal de 5%, de testes "quasi-*t*" que usam as estimativas das variâncias calculadas a partir dos estimadores aqui considerados. O interesse recai em testar a hipótese H_0: $\beta_2 = 1$ contra H_1: $\beta_2 \neq 1$ contra. Como podemos observar a partir da análise desta tabela, o teste baseado no estimador MQO torna-se mais liberal à medida que aumenta o grau de heteroscedasticidade. No entanto, as maiores distorções de tamanho estão associadas ao estimador HC11 para todos os tamanhos amostrais e valores de λ, as taxas de rejeição dos testes baseados neste estimador são em média próximas a 40%. É interessante notar que os estimadores que se mostraram mais viesados (HC13 a HC18) agora apresentam as taxas de rejeição dos testes neles baseados, mais próximas do que seria esperado de acordo com o nível nominal, ou seja, comportam-se melhor que todos os outros HCCMEs, sendo que, para amostras de tamanho moderado, o estimador HC15 se destaca dos demais.

Tabela 1 – Viés relativo total, geração sob distribuição uniforme (sem pontos de alto alavancagem)

n	γ	λ	MQO	HC0	HC1	HC2	HC3	HC4	HC5	HC6	HC7	HC8	HC9	HC10	HC11	HC12	HC13	HC14	HC15	HC16	HC17	HC18
50	0,0	1,0	0,0001	0,1032	0,0241	0,0048	0,1195	0,0566	0,0566	0,0566	0,0566	0,0566	0,0566	0,0280	1,5916	0,0613	0,1794	0,2412	0,3049	0,3706	0,4383	0,5080
	1,0	2,7	0,3574	0,0970	0,0177	0,0098	0,1232	0,0600	0,0600	0,0600	0,0600	0,0600	0,0600	0,0423	1,5790	0,0557	0,1825	0,2436	0,3066	0,3715	0,4383	0,5072
	2,0	7,1	0,8449	0,1005	0,0214	0,0112	0,1196	0,0569	0,0569	0,0569	0,0569	0,0569	0,0569	0,0629	1,5794	0,0621	0,1789	0,2400	0,3030	0,3679	0,4347	0,5037
	3,0	18,9	1,3656	0,0960	0,0220	0,0191	0,1294	0,0678	0,0678	0,0678	0,0678	0,0678	0,0678	0,0633	1,5918	0,0704	0,1902	0,2528	0,3174	0,3840	0,4527	0,5236
	4,0	50,4	1,7379	0,1235	0,0454	0,0284	0,1053	0,0461	0,0461	0,0461	0,0461	0,0461	0,0461	0,0792	1,6183	0,0454	0,1671	0,2308	0,2966	0,3645	0,4345	0,5067
	4,5	82,3	1,8585	0,1220	0,0437	0,0269	0,1112	0,0528	0,0528	0,0528	0,0528	0,0528	0,0528	0,0653	1,6308	0,0501	0,1742	0,2392	0,3063	0,3756	0,4471	0,5209
100	0,0	1,0	0,0006	0,0543	0,0146	0,0004	0,0550	0,0247	0,0247	0,0247	0,0247	0,0247	0,0247	0,0223	1,5748	0,0271	0,0834	0,1121	0,1413	0,1710	0,2011	0,2316
	1,0	2,7	0,3483	0,0510	0,0112	0,0022	0,0570	0,0264	0,0264	0,0264	0,0264	0,0264	0,0264	0,0306	1,5629	0,0294	0,0850	0,1134	0,1423	0,1716	0,2013	0,2314
	2,0	7,1	0,8444	0,0487	0,0089	0,0069	0,0596	0,0292	0,0292	0,0292	0,0292	0,0292	0,0292	0,0401	1,5621	0,0319	0,0877	0,1162	0,1451	0,1744	0,2042	0,2345
	3,0	18,9	1,3640	0,0492	0,0093	0,0071	0,0614	0,0316	0,0316	0,0316	0,0316	0,0316	0,0316	0,0475	1,5760	0,0331	0,0900	0,1191	0,1487	0,1787	0,2091	0,2400
	4,0	50,4	1,7530	0,0566	0,0169	0,0113	0,0568	0,0281	0,0281	0,0281	0,0281	0,0281	0,0281	0,0521	1,5997	0,0278	0,0863	0,1162	0,1466	0,1774	0,2088	0,2406
	4,5	82,3	1,8803	0,0471	0,0126	0,0121	0,0690	0,0407	0,0407	0,0407	0,0407	0,0407	0,0407	0,0367	1,6117	0,0393	0,0992	0,1299	0,1610	0,1926	0,2247	0,2573
150	0,0	1,0	0,0029	0,0354	0,0089	0,0009	0,0371	0,0171	0,0171	0,0171	0,01 711	0,0171	0,0171	0,0158	1,5691	0,0187	0,0557	0,0745	0,0935	0,1126	0,1320	0,1515
	1,0	2,7	0,3544	0,0301	0,0035	0,0054	0,0416	0,0213	0,0213	0,0213	0,0213	0,0213	0,0213	0,0189	1,5553	0,0234	0,0599	0,0785	0,0972	0,1161	0,1352	0,1544
	2,0	7,1	0,8441	0,0371	0,0106	0,0042	0,0344	0,0144	0,0144	0,0144	0,0144	0,0144	0,0144	0,0349	1,5571	0,0163	0,0528	0,0713	0,0900	0,1089	0,1279	0,1472
	3,0	18,9	1,3640	0,0343	0,0078	0,0059	0,0388	0,0192	0,0192	0,0192	0,0192	0,0192	0,0192	0,0385	1,5703	0,0202	0,0575	0,0764	0,0956	0,1149	0,1344	0,1540
	4,0	50,4	1,7544	0,0357	0,0092	0,0066	0,0397	0,0207	0,0207	0,0207	0,0207	0,0207	0,0207	0,0394	1,5941	0,0206	0,0590	0,0786	0,0983	0,1182	0,1384	0,1587
	4,5	82,3	1,8728	0,0406	0,0141	0,0093	0,0359	0,0173	0,0173	0,0173	0,0173	0,0173	0,0173	0,0412	1,6079	0,0165	0,0555	0,0754	0,0954	0,1156	0,1361	0,1567
200	0,0	1,0	0,0005	0,0276	0,0077	0,0007	0,0267	0,0117	0,0117	0,0117	0,0117	0,0117	0,0117	0,0138	1,5663	0,0129	0,0405	0,0544	0,0684	0,0825	0,0968	0,1111
	1,0	2,7	0,3489	0,0257	0,0058	0,0017	0,0278	0,0127	0,0127	0,0127	0,0127	0,0127	0,0127	0,0185	1,5537	0,0143	0,0414	0,0551	0,0690	0,0829	0,0969	0,1110
	2,0	7,1	0,8433	0,0264	0,0065	0,0031	0,0273	0,0123	0,0123	0,0123	0,0123	0,0123	0,0123	0,0264	1,5542	0,0137	0,0409	0,0547	0,0685	0,0825	0,0965	0,1107
	3,0	18,9	1,3646	0,0283	0,0084	0,0046	0,0263	0,0117	0,0117	0,0117	0,0117	0,0117	0,0117	0,0340	1,5677	0,0125	0,0402	0,0542	0,0683	0,0826	0,0969	0,1113
	4,0	50,4	1,7558	0,0252	0,0056	0,0051	0,0313	0,0171	0,0171	0,0171	0,0171	0,0171	0,0171	0,0315	1,5909	0,0170	0,0456	0,0801	0,0747	0,0895	0,1043	0,1192
	4,5	82,3	1,8789	0,0280	0,0061	0,0056	0,0314	0,0175	0,0175	0,0175	0,0175	0,0175	0,0175	0,0303	1,6041	0,0168	0,0460	0,0607	0,0756	0,0905	0,1056	0,1208

Tabela 2 – EQM total, geração sob distribuição uniforme (sem pontos de alto alavancagem)

n	γ	λ	MQO	HC0	HC1	HC2	HC3	HC4	HC5	HC6	HC7	HC8	HC9	HC10	HC11	HC12	HC13	HC14	HC15	HC16	HC17	HC18
50	0,0	1,0	0,0023	0,0038	0,0040	0,0041	0,0049	0,0045	0,0045	0,0045	0,0045	0,0045	0,0045	0,0041	0,0352	0,0045	0,0054	0,0061	0,0069	0,0079	0,0091	0,0104
	1,0	2,7	0,0624	0,0862	0,0904	0,0942	0,1109	0,1022	0,1022	0,1022	0,1022	0,1022	0,1022	0,0939	0,7240	0,1014	0,1230	0,1379	0,1557	0,1769	0,2015	0,2300
	2,0	7,1	2,3375	3,2158	3,3897	3,5362	4,1135	3,8343	3,8343	3,8343	3,8343	3,8343	3,8343	3,5911	20,3469	3,7900	4,5128	4,9945	5,5655	6,2334	7,0063	7,8829
	3,0	18,9	105,8260	142,8740	151,4650	158,5530	184,0577	172,7616	172,7616	172,7616	172,7616	172,7616	172,7616	163,8270	714,0866	169,9710	201,0470	221,1900	244,7570	272,0370	303,3410	339,0030
	4,0	50,4	5411,8	6723,3	7094,1	7422,6	8529,1	8083,1	8083,1	8083,1	8083,1	8083,1	8083,1	7778,2	28744,5	7920,5	9258,4	10118,6	11121,3	12278,6	13603,4	15109,9
	4,5	82,3	39152,1	45796,4	48414,1	50702,9	58343,0	55298,0	55298,0	55298,0	55298,0	55298,0	55298,0	53388,4	187241,9	54151,2	63343,9	69225,0	76062,3	83937,8	92939,5	103162,0
100	0,0	1,0	0,0003	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005	0,0086	0,0005	0,0008	0,0006	0,0007	0,0007	0,0008	0,0008
	1,0	2,7	0,0085	0,0110	0,0113	0,0115	0,0125	0,0120	0,0120	0,0120	0,0120	0,0120	0,0120	0,0115	0,1769	0,0119	0,0132	0,0140	0,0149	0,0160	0,0172	0,0186
	2,0	7,1	0,3724	0,4184	0,4297	0,4388	0,4730	0,4588	0,4588	0,4588	0,4588	0,4588	0,4588	0,4385	4,9646	0,4541	0,4957	0,5223	0,5530	0,5879	0,6273	0,6714
	3,0	18,9	18,8083	18,2932	18,8386	19,2699	20,7583	20,1080	20,1080	20,1080	20,1080	20,1080	20,1080	19,3999	174,2151	19,9473	21,7085	22,8041	24,0515	25,4570	27,0275	28,7699
	4,0	50,4	1034,76	849,81	874,24	894,11	959,56	932,71	932,71	932,71	932,71	932,71	932,71	906,78	6987,30	924,07	1000,82	1048,09	1101,66	1161,78	1228,74	1302,83
	4,5	82,3	7751,3	6236,9	6438,3	6598,4	7105,4	6907,4	6907,4	6907,4	6907,4	6907,4	6907,4	6716,8	45496,6	6832,9	7417,5	7771,1	8167,8	8609,7	9098,6	9636,5
150	0,0	1,0	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0001	0,0038	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
	1,0	2,7	0,0028	0,0032	0,0033	0,0034	0,0035	0,0034	0,0034	0,0034	0,0034	0,0034	0,0034	0,0033	0,0779	0,0034	0,0037	0,0038	0,0040	0,0042	0,0045	0,0047
	2,0	7,1	0,1335	0,1223	0,1242	0,1259	0,1321	0,1292	0,1292	0,1292	0,1292	0,1292	0,1292	0,1257	2,1895	0,1287	0,1382	0,1410	0,1465	0,1528	0,1598	0,1676
	3,0	18,9	7,2266	5,5796	5,6863	5,7706	6,0560	5,9325	5,9325	5,9325	5,9325	5,9325	5,9325	5,7806	76,8090	5,9012	6,2359	6,4416	6,6737	6,9331	7,2205	7,5368
	4,0	50,4	417,789	264,341	269,645	273,856	287,322	281,931	281,931	281,931	281,931	281,931	281,931	275,624	3077,716	280,080	295,610	304,977	315,453	327,071	339,864	353,867
	4,5	82,3	3189,26	1887,15	1923,31	1952,92	2044,86	2009,82	2009,82	2009,82	2009,82	2009,82	2009,82	1970,29	20095,80	1995,52	2101,11	2164,51	2235,26	2313,58	2399,67	2493,78
200	0,0	1,0	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0021	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	1,0	2,7	0,0013	0,0014	0,0014	0,0014	0,0015	0,0014	0,0014	0,0014	0,0014	0,0014	0,0014	0,0014	0,0437	0,0014	0,0015	0,0015	0,0016	0,0017	0,0017	0,0018
	2,0	7,1	0,0672	0,0523	0,0529	0,0535	0,0555	0,0545	0,0545	0,0545	0,0545	0,0545	0,0545	0,0533	1,2265	0,0544	0,0568	0,0583	0,0600	0,0620	0,0642	0,0666
	3,0	18,9	3,7997	2,3987	2,4313	2,4572	2,5446	2,5069	2,5069	2,5069	2,5069	2,5069	2,5069	2,4576	43,0407	2,4972	2,5995	2,6620	2,7325	2,8110	2,8976	2,9927
	4,0	50,4	221,56	110,45	112,14	113,45	117,66	115,96	115,96	115,96	115,96	115,96	115,96	113,82	1723,32	115,40	120,23	123,13	126,35	129,91	133,82	138,07
	4,5	82,3	1696,58	791,94	804,23	813,85	843,77	832,17	832,17	832,17	832,17	832,17	832,17	817,79	11235,80	827,79	861,85	882,08	904,49	929,13	956,07	985,33

Tabela 3 – Tamanhos estimados do teste "quase-t" com $\alpha = 5\%$ e geração sob distribuição uniforme (sem pontos de alto alavancagem)

n	γ	λ	MQO	HCO	HC1	HC2	HC3	HC4	HC5	HC6	HC7	HC8	HC9	HC10	HC11	HC12	HC13	HC14	HC15	HC16	HC17	HC18
50	0,0	1,0	5,85	6,44	5,95	5,80	5,11	5,55	5,55	5,55	5,55	5,55	5,55	5,98	38,98	5,51	4,85	4,50	4,28	4,03	3,87	3,69
	1,0	2,7	5,83	6,84	6,28	6,10	5,45	5,80	5,80	5,80	5,80	5,80	5,80	6,27	39,13	5,77	5,16	4,83	4,62	4,32	4,00	3,73
	2,0	7,1	6,29	6,82	6,31	6,18	5,57	5,87	5,87	5,87	5,87	5,87	5,87	6,26	39,93	5,86	5,31	4,99	4,68	4,38	4,15	3,93
	3,0	18,9	7,25	7,09	6,53	6,34	5,61	6,03	6,03	6,03	6,03	6,03	6,03	6,45	41,07	5,93	5,27	4,94	4,70	4,38	4,13	3,92
	4,0	50,4	9,13	7,75	7,09	6,85	6,20	6,57	6,57	6,57	6,57	6,57	6,57	6,96	42,06	6,53	5,79	5,38	4,96	4,57	4,29	4,04
	4,5	82,3	9,72	7,56	7,15	6,93	6,21	6,57	6,57	6,57	6,57	6,57	6,57	6,92	42,51	6,55	5,74	5,35	4,97	4,63	4,35	4,00
100	0,0	1,0	5,17	5,78	5,53	5,45	5,06	5,20	5,20	5,20	5,20	5,20	5,20	5,56	38,32	5,18	4,96	4,81	4,64	4,47	4,34	4,23
	1,0	2,7	5,25	6,08	5,79	5,72	5,42	5,59	5,59	5,59	5,59	5,59	5,59	5,84	38,23	5,58	5,19	5,05	4,80	4,70	4,55	4,44
	2,0	7,1	5,87	5,90	5,63	5,56	5,30	5,48	5,48	5,48	5,48	5,48	5,48	5,62	38,50	5,46	5,12	4,97	4,81	4,60	4,49	4,41
	3,0	18,9	6,84	5,97	5,80	5,74	5,37	5,60	5,60	5,60	5,60	5,60	5,80	5,80	38,62	5,59	5,28	5,12	5,01	4,83	4,66	4,51
	4,0	50,4	8,64	6,29	6,04	5,95	5,55	5,71	5,71	5,71	5,71	5,71	5,71	5,98	40,52	5,72	5,40	5,22	5,07	4,90	4,83	4,69
	4,5	82,3	9,25	6,36	6,17	6,00	5,64	5,84	5,84	5,84	5,84	5,84	5,84	6,04	41,15	5,83	5,36	5,14	5,01	4,85	4,72	4,55
150	0,0	1,0	5,57	5,86	5,67	5,65	5,46	5,58	5,58	5,58	5,58	5,58	5,58	5,73	38,30	5,58	5,29	5,22	5,18	5,04	4,96	4,85
	1,0	2,7	5,02	5,51	5,33	5,28	4,98	5,15	5,15	5,15	5,15	5,15	5,15	5,35	37,80	5,14	4,79	4,74	4,65	4,57	4,47	4,35
	2,0	7,1	5,89	5,57	5,44	5,40	5,23	5,32	5,32	5,32	5,32	5,32	5,32	5,47	37,83	5,31	5,14	5,02	4,94	4,85	4,76	4,68
	3,0	18,9	7,27	6,00	5,87	5,78	5,56	5,70	5,70	5,70	5,70	5,70	5,70	5,85	39,13	5,70	5,47	5,34	5,18	5,04	4,88	4,82
	4,0	50,4	8,85	6,02	5,85	5,80	5,57	5,67	5,67	5,67	5,67	5,67	5,67	5,82	40,28	5,66	5,35	5,24	5,13	4,94	4,84	4,75
	4,5	82,3	9,61	6,31	6,10	5,99	5,77	5,87	5,87	5,87	5,87	5,87	5,87	6,03	40,23	5,90	5,68	5,55	5,44	5,29	5,25	5,13
200	0,0	1,0	5,14	5,44	5,32	5,30	5,09	5,18	5,18	5,18	5,18	5,18	5,18	5,36	38,32	5,18	5,03	4,95	4,87	4,76	4,69	4,62
	1,0	2,7	5,19	5,48	5,39	5,36	5,22	5,32	5,32	5,32	5,32	5,32	5,32	5,44	37,39	5,32	5,10	5,04	4,97	4,92	4,82	4,73
	2,0	7,1	5,98	5,59	5,44	5,39	5,27	5,32	5,32	5,32	5,32	5,32	5,32	5,50	38,08	5,31	5,17	5,08	5,01	4,96	4,85	4,75
	3,0	18,9	7,26	5,92	5,86	5,83	5,70	5,77	5,77	5,77	5,77	5,77	5,77	5,86	38,39	5,76	5,57	5,49	5,38	5,28	5,14	5,06
	4,0	50,4	8,51	5,67	5,47	5,44	5,25	5,38	5,38	5,38	5,38	5,38	5,38	5,46	39,08	5,38	5,17	5,06	4,91	4,82	4,73	4,69
	4,5	82,3	9,45	6,02	5,92	5,88	5,73	5,82	5,82	5,82	5,82	5,82	5,82	5,89	40,19	5,82	5,66	5,62	5,52	5,39	5,27	5,18

Outra percepção importante é que para um grau de heterocedasticidade severo, $\gamma = 4.5$, o que corresponde a $\lambda = 82.34$, o desempenho destes estimadores é bastante satisfatório e superior ao desempenho do estimador HC4. Por exemplo, para $n = 100$ e $\lambda = 82.34$, as taxas de rejeição dos testes baseados nos estimadores HC4 e HC12 a HC18 são, respectivamente, 5.84% , 5.83%, 5.36%, 5.14%, 4.85%, 4.72% e 4.55%. Os testes baseados nos estimadores HC4 a HC9 são menos liberais que os testes baseados nos estimadores HC1 e HC2, no entanto, o estimador HC3 empregado em testes de hipótese fornece resultados mais confiáveis que os resultados associados ao estimador HC4.

As Tabelas de 4 até 6 sumarizam os resultados da investigação considerando a distribuição exponencial para x. Neste caso, em todos tamanhos de amostras o maior valor de h_i corresponde a aproximadamente duas vezes o limiar crítico $3p/n$. Os dados dispostos na Tabela 4 se referem ao viés relativo total dos diferentes estimadores e revelam que, sob heteroscedasticidade, tipicamente, o estimador HC12 apresenta os menores vieses relativos totais entre todos os HCCMEs avaliados. Quando o expoente fixo δ_J da quantidade $(1-h_i)$ é maior ou igual a dois e meio (2.5), os estimadores HC13 até HC18 se tornam consideravelmente viesados. Já, entre os estimadores HCJ que consideram δ_{J_i}, $J = 13,...,17$, $i = 1,...,n$,o melhor desempenho para todos tamanhos de amostra e valores de λ é o do estimador HC5. Por exemplo, para $n = 200$ e $\lambda = 78.28$, os vieses relativos totais dos estimadores HC4 a HC9 são, respectivamente, 29.75%, 10.85% , 17.20%, 23.43%, 36.32% e 43.14%. Na Tabela 5, estão apresentados os erros quadráticos médios dos estimadores avaliados. A análise desta tabela evidencia que, em relação ao EQM, entre os estimadores HC4 até HC9, é o estimador HC5 que apresenta o melhor desempenho. No entanto, no geral, o melhor desempenho, no que diz respeito a este critério, é do estimador HC0. Os desempenhos dos estimadores HC10 e HC12 são semelhantes e satisfatórios; por exemplo, para $n = 100$ e $\lambda = 26.31$, os erros quadráticos médios destes estimadores são, respectivamente, 0.01447 e 0.01429. Quanto ao resultado das taxas de rejeição dos teste "quasi-t" baseados nos HCCMEs avaliados, Tabela 6, os estimadores utilizados nos testes que apresentaram os melhores desempenhos, inclusive superior ao HC4, são os HC8 e HC9. Com relação aos estimadores que utilizam expoente fixo para a quantidade $(1-h_i)$, quanto maior o expoente menores são as taxas de rejeição dos testes associados; no entanto, à medida que cresce o tamanho da amostra as taxas de rejeição tendem a aumentar.

Tabela 4 – Viés relativo total, geração sob distribuição exponencial (com pontos de alto alavancagem)

n	γ	λ	MQO	HC0	HC1	HC2	HC3	HC4	HC5	HC6	HC7	HC8	HC9	HC10	HC11	HC12	HC13	HC14	HC15	HC16	HC17	HC18
50	0,0	1,0	0,0003	0,1837	0,1081	0,0035	0,2252	0,7007	0,2904	0,4202	0,5524	0,8769	1,0862	0,0878	0,6347	0,1043	0,3621	0,5178	0,6955	0,8991	1,1330	1,4028
	0,2	3,0	0,3834	0,2851	0,2136	0,0627	0,2255	0,9314	0,3445	0,5256	0,7158	1,1876	1,4918	0,0982	0,8023	0,0718	0,4018	0,6046	0,8385	1,1091	1,4229	1,7878
	0,4	8,8	0,6495	0,4166	0,3506	0,1423	0,2222	1,2268	0,4079	0,6543	0,9209	1,5901	2,0215	0,1159	1,0129	0,0268	0,4487	0,7117	1,0180	1,3752	1,7926	2,2812
	0,6	26,3	0,9433	0,5470	0,4865	0,2164	0,2315	1,5660	0,4879	0,8058	1,1581	2,0504	2,6257	0,1077	1,2274	0,0096	0,5130	0,8425	1,2286	1,6816	2,2137	2,8396
	0,8	78,3	1,2096	0,6805	0,6255	0,3147	0,1865	1,7463	0,4904	0,8559	1,2674	2,3151	2,9906	0,0756	1,4071	0,0840	0,5039	0,8769	1,3157	1,8324	2,4412	3,1592
100	0,0	1,0	0,0007	0,0917	0,0527	0,0014	0,0988	0,2439	0,1179	0,1630	0,2034	0,2876	0,3349	0,0396	0,5760	0,0474	0,1532	0,2106	0,2713	0,3355	0,4036	0,4758
	0,2	3,0	0,3844	0,1370	0,0990	0,0190	0,1142	0,3545	0,1592	0,2266	0,2899	0,4243	0,4999	0,0658	0,7301	0,0456	0,1871	0,2647	0,3474	0,4354	0,5293	0,6295
	0,4	8,8	0,6246	0,2006	0,1639	0,0492	0,1237	0,4816	0,1994	0,2938	0,3859	0,5851	0,6970	0,0860	0,9200	0,0344	0,2192	0,3214	0,4307	0,5479	0,6734	0,8079
	0,6	26,3	0,9129	0,2692	0,2338	0,0821	0,1332	0,6186	0,2409	0,3643	0,4883	0,7595	0,9120	0,0855	1,1205	0,0218	0,2529	0,3815	0,5197	0,6682	0,8278	0,9995
	0,8	78,3	1,1686	0,3553	0,3217	0,1454	0,0975	0,6695	0,2267	0,3692	0,5152	0,8364	1,0169	0,0781	1,2870	0,0284	0,2330	0,3789	0,5359	0,7051	0,8872	1,0835
150	0,0	1,0	0,0029	0,0641	0,0379	0,0040	0,0602	0,1433	0,0707	0,0977	0,1209	0,1669	0,1917	0,0279	0,5584	0,0275	0,0940	0,1291	0,1654	0,2030	0,2420	0,2824
	0,2	3,0	0,3860	0,0976	0,0719	0,0187	0,0666	0,2065	0,0933	0,1337	0,1703	0,2447	0,2848	0,0493	0,7073	0,0231	0,1118	0,1588	0,2078	0,2588	0,3120	0,3673
	0,4	8,8	0,6205	0,1421	0,1170	0,0399	0,0715	0,2830	0,1178	0,1748	0,2287	0,3401	0,4002	0,0747	0,8919	0,0146	0,1309	0,1929	0,2577	0,3255	0,3963	0,4703
	0,6	26,3	0,9018	0,1962	0,1718	0,0702	0,0678	0,3532	0,1337	0,2077	0,2798	0,4305	0,5119	0,0903	1,0856	0,0028	0,1417	0,2191	0,3001	0,3850	0,4740	0,5672
	0,8	78,3	1,1526	0,2351	0,2113	0,0893	0,0711	0,4194	0,1531	0,2414	0,3292	0,5144	0,6144	0,0721	1,2399	0,0110	0,1573	0,2476	0,3424	0,4418	0,5462	0,6557
200	0,0	1,0	0,0004	0,0421	0,0223	0,0035	0,0514	0,1107	0,0588	0,0784	0,0949	0,1271	0,1442	0,0225	0,5461	0,0271	0,0763	0,1018	0,1280	0,1549	0,1825	0,2109
	0,2	3,0	0,3858	0,0663	0,0468	0,0061	0,0577	0,1584	0,0772	0,1066	0,1328	0,1850	0,2126	0,0380	0,6932	0,0253	0,0910	0,1253	0,1607	0,1971	0,2346	0,2733
	0,4	8,8	0,6167	0,0981	0,0789	0,0198	0,0637	0,2164	0,0977	0,1392	0,1779	0,2563	0,2977	0,0562	0,8737	0,0213	0,1075	0,1527	0,1994	0,2477	0,2975	0,3490
	0,6	26,3	0,8938	0,1302	0,1113	0,0328	0,0715	0,2794	0,1204	0,1747	0,2269	0,3339	0,3906	0,0617	1,0603	0,0184	0,1263	0,1831	0,2418	0,3026	0,3655	0,4307
	0,8	78,3	1,1461	0,1813	0,1630	0,0703	0,0488	0,2976	0,1085	0,1720	0,2344	0,3633	0,4314	0,0731	1,2179	0,0118	0,1116	0,1766	0,2440	0,3139	0,3862	0,4612

Tabela 5 EQM total, geração sob distribuição exponencial (com pontos de alto alavancagem)

n	γ	λ	MQO	HCO	HC1	HC2	HC3	HC4	HC5	HC6	HC7	HC8	HC9	HC10	HC11	HC12	HC13	HC14	HC15	HC16	HC17	HC18
50	0,0	1,0	0,0001	0,0001	0,0001	0,0001	0,0002	0,0005	0,0002	0,0003	0,0004	0,0007	0,0010	0,0002	0,0002	0,0002	0,0003	0,0003	0,0005	0,0006	0,0009	0,0012
	0,2	3,0	0,0003	0,0005	0,0005	0,0007	0,0013	0,0047	0,0017	0,0024	0,0034	0,0066	0,0094	0,0010	0,0007	0,0009	0,0018	0,0025	0,0036	0,0051	0,0074	0,0105
	0,4	8,8	0,0042	0,0045	0,0046	0,0066	0,0121	0,0493	0,0170	0,0243	0,0346	0,0703	0,1005	0,0097	0,0058	0,0087	0,0172	0,0248	0,0358	0,0519	0,0750	0,1080
	0,6	26,3	0,0546	0,0509	0,0522	0,0761	0,1424	0,6012	0,2026	0,2920	0,4191	0,8633	1,2392	0,1213	0,0589	0,1021	0,2035	0,2943	0,4275	0,6209	0,8999	1,3000
	0,8	78,3	0,6985	0,6054	0,6223	0,8965	1,6628	7,0681	2,3657	3,4108	4,9108	10,1809	14,6546	1,4664	0,6748	1,1961	2,3710	3,4259	4,9772	7,2362	10,5021	15,1971
100	0,0	1,0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	0,2	3,0	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0001	0,0002	0,0002	0,0003	0,0003	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004
	0,4	8,8	0,0009	0,0008	0,0008	0,0010	0,0013	0,0026	0,0016	0,0019	0,0022	0,0032	0,0038	0,0011	0,0012	0,0011	0,0016	0,0019	0,0023	0,0027	0,0033	0,0040
	0,6	26,3	0,0129	0,0098	0,0100	0,0124	0,0168	0,0338	0,0199	0,0237	0,0283	0,0405	0,0486	0,0145	0,0123	0,0143	0,0199	0,0238	0,0286	0,0346	0,0417	0,0505
	0,8	78,3	0,1665	0,1199	0,1226	0,1505	0,2027	0,4070	0,2395	0,2851	0,3403	0,4879	0,5856	0,1807	0,1423	0,1734	0,2397	0,2857	0,3426	0,4124	0,4975	0,6006
150	0,0	1,0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	0,2	3,0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0000	0,0000	0,0000	0,0001	0,0001	0,0000	0,0001	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001
	0,4	8,8	0,0004	0,0003	0,0003	0,0003	0,0004	0,0006	0,0004	0,0005	0,0005	0,0006	0,0007	0,0003	0,0005	0,0003	0,0004	0,0005	0,0005	0,0006	0,0007	0,0008
	0,6	26,3	0,0056	0,0031	0,0031	0,0036	0,0044	0,0069	0,0049	0,0055	0,0061	0,0078	0,0088	0,0039	0,0050	0,0039	0,0049	0,0055	0,0062	0,0070	0,0080	0,0091
	0,8	78,3	0,0725	0,0409	0,0416	0,0478	0,0583	0,0928	0,0650	0,0730	0,0822	0,1049	0,1189	0,0529	0,0581	0,0526	0,0651	0,0731	0,0826	0,0936	0,1064	0,1212
200	0,0	1,0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	0,2	3,0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	0,4	8,8	0,0002	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0003	0,0001	0,0003	0,0001	0,0002	0,0002	0,0002	0,0002	0,0002	0,0003
	0,6	26,3	0,0031	0,0014	0,0014	0,0016	0,0019	0,0027	0,0020	0,0022	0,0024	0,0029	0,0032	0,0017	0,0027	0,0017	0,0020	0,0022	0,0024	0,0027	0,0030	0,0033
	0,8	78,3	0,0403	0,0179	0,0182	0,0201	0,0233	0,0330	0,0253	0,0275	0,0301	0,0362	0,0399	0,0215	0,0312	0,0216	0,0253	0,0276	0,0302	0,0332	0,0366	0,0405

Tabela 6 – Tamanhos estimados do teste "quase-t" com $\alpha = 5\%$ e geração sob distribuição exponencial

n	γ	λ	MQO	HC0	HC1	HC2	HC3	HC4	HC5	HC6	HC7	HC8	HC9	HC10	HC11	HC12	HC13	HC14	HC15	HC16	HC17	HC18
50	0,0	1,0	5,38	9,15	8,59	7,58	6,31	4,65	6,07	5,42	5,06	4,35	4,02	7,58	15,49	6,83	5,69	5,14	4,36	3,81	3,32	2,81
	0,2	3,0	12,16	11,60	11,01	9,55	7,74	5,59	7,18	6,45	6,01	5,31	4,97	9,28	20,53	8,55	6,89	6,16	5,45	4,91	4,44	3,92
	0,4	8,8	22,83	13,37	12,64	10,58	8,50	5,86	7,76	6,96	6,34	5,37	5,05	9,89	25,95	9,37	7,56	6,69	5,80	5,06	4,59	4,07
	0,6	26,3	35,09	14,83	14,00	11,63	9,06	5,68	8,00	6,91	6,05	5,11	4,55	10,34	31,07	10,25	7,81	6,64	5,69	4,92	4,26	3,53
	0,8	78,3	46,87	16,09	15,18	12,23	9,17	5,57	8,03	6,84	6,11	5,08	4,64	10,51	36,08	10,53	7,92	6,65	5,79	4,93	4,19	3,58
100	0,0	1,0	5,02	7,28	7,01	6,40	5,64	4,73	5,40	5,12	4,94	4,55	4,39	6,41	13,78	6,00	5,29	4,97	4,71	4,31	4,00	3,68
	0,2	3,0	12,26	8,52	8,20	7,53	6,46	5,40	6,21	5,80	5,57	5,06	4,75	7,27	17,83	7,00	6,05	5,66	5,34	4,99	4,56	4,27
	0,4	8,8	21,82	9,39	8,99	8,08	6,85	5,23	6,35	5,82	5,52	4,97	4,71	7,68	20,94	7,43	6,27	5,75	5,31	4,89	4,49	4,06
	0,6	26,3	33,18	10,42	10,08	8,74	7,31	5,31	6,73	6,21	5,70	4,93	4,52	8,28	25,26	8,03	6,67	6,11	5,53	5,04	4,41	3,87
	0,8	78,3	43,61	11,10	10,66	9,21	7,69	5,46	6,95	6,34	5,89	4,94	4,56	8,45	28,82	8,36	6,89	6,29	5,71	5,20	4,63	4,18
150	0,0	1,0	5,29	6,54	6,26	5,98	5,51	5,04	5,39	5,27	5,14	4,94	4,80	5,99	13,76	5,76	5,32	5,18	4,92	4,75	4,54	4,36
	0,2	3,0	12,02	7,59	7,39	6,96	6,42	5,64	6,20	6,02	5,84	5,46	5,32	6,87	16,53	6,65	6,17	5,93	5,69	5,41	5,18	4,97
	0,4	8,8	21,82	7,92	7,75	7,09	6,42	5,17	6,01	5,67	5,42	4,91	4,72	6,93	20,06	6,68	5,93	5,57	5,31	4,91	4,70	4,37
	0,6	26,3	33,25	8,34	8,14	7,37	6,52	5,34	6,30	6,07	5,69	5,07	4,85	7,07	23,69	6,98	6,29	5,98	5,53	5,13	4,78	4,46
	0,8	78,3	43,33	9,16	8,96	7,95	6,71	5,13	6,25	5,82	5,46	4,77	4,45	7,52	26,79	7,37	6,24	5,78	5,37	4,91	4,51	4,09
200	0,0	1,0	5,58	6,46	6,41	6,18	5,77	5,29	5,68	5,53	5,39	5,23	5,05	6,15	12,94	5,94	5,60	5,42	5,25	5,11	4,88	4,74
	0,2	3,0	12,02	6,92	6,80	6,49	5,96	5,27	5,83	5,57	5,38	5,11	5,00	6,42	16,35	6,24	5,73	5,46	5,29	5,11	4,90	4,69
	0,4	8,8	22,47	7,59	7,41	6,90	6,38	5,48	6,15	5,95	5,78	5,27	5,04	6,79	19,63	6,64	6,12	5,90	5,62	5,31	4,99	4,74
	0,6	26,3	32,94	7,73	7,52	7,02	6,43	5,28	6,09	5,80	5,60	5,02	4,79	6,86	22,49	6,73	6,06	5,76	5,47	5,12	4,81	4,54
	0,8	78,3	43,38	7,99	7,82	7,11	6,42	5,18	5,99	5,60	5,41	4,92	4,62	6,93	25,63	6,79	5,97	5,59	5,35	5,07	4,74	4,46

5. Conclusões

Na seção 4 estão apresentados resultados referentes à avaliação das propriedades em amostras finitas de diferentes estimadores consistentes para a matriz de covariâncias do estimador de MQO. Foram utilizados como critérios de avaliação o viés relativo total, erro quadrático médio e comportamento de estatísticas "quasi-t" construídas a partir dos diferentes HCCMEs. As investigações consideraram diferentes níveis de heteroscedasticidade e tamanhos amostrais. Os resultados obtidos revelam que, para dados sem pontos de alavanca, o estimador HC2 apresenta o menor viés relativo total, seguido pelo estimador HC1 e, dentre as novas propostas, pelo estimador HC12 seguido do estimador HC10. No entanto, quanto ao EQM, em geral, o estimador de White é o que apresenta o melhor desempenho entre todos os HCCMEs e entre as novas propostas o estimador HC10 é o que se comporta melhor. Os estimadores HC4 a HC9 apresentaram resultados idênticos quanto a todos os critérios investigados, o que é razoável, pois neste caso não existem pontos que exerçam influência desproporcional sobre a estimação do modelo. Quanto aos estimadores que utilizam expoentes fixos para a quantidade $(1-h_i)$, HC13 a HC18, à medida que o expoente cresce o desempenho piora, tanto em relação ao viés relativo total quanto em relação ao erro quadrático médio; no entanto, tornam-se mais confiáveis no que concerne inferência "quasi-t", com destaque especial para o estimador HC15. Já, quando a investigação considera a existência de pontos de alta alavancagem, o novo estimador HC12 apresenta os menores vieses relativos totais e o estimador HC0, o melhor desempenho quanto ao EQM. Novamente, os estimadores HC13 a HC18 são consideravelmente viesados, contudo, não são os melhores quanto às taxas de rejeição de testes associados. Entre os estimadores HC4 a HC9, é o estimador HC5 que se comporta melhor, tanto em relação ao erro quadrático médio quanto em relação ao viés relativo total. No entanto, no que concerne as taxas de rejeição dos testes "quasi-t" baseados nos diferentes estimadores, HC8 e HC9, que em geral são mais viesados que o estimador HC4, são os mais confiáveis e seus desempenhos estão entre os melhores quando há um ponto de alavanca muito forte, em particular, quando utilizamos o estimador HC9.

Referências bibliográficas

BELSLEY, D. A. (2002). An investigation of an unbiased correction for heteroskedasticity and the effects of misspecifying the skedastic function. *Journal of Economic Dynamics and Control*, 26, 1379-1396.

BREUSCH, T.S. & PAGAN, A.R. (1979). A simple test for heteroskedasticity and random coefficient variation. *Econometrica*, 47, 1287-1294.

CRIBARI-NETO, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics and Data Analysis*, 45, 215-233.

CRIBARI-NETO, F., FERRARI, S.L.P. & CORDEIRO, G.M. (2000). Improved heterocedasticity-consistent covariance matrix estimators. *Biometrika*, 87, 907-918.

CRIBARI-NETO, F. & ZARKOS, S.G. (1999). Bootstrap methods for heteroskedastic regression models: evidence on estimation and testing. *Econometric Reviews*, 18, 211-28.

BELSLEY, D. A. (2002). An investigation of an unbiased correction for heteroskedasticity and the effects of misspecifying the skedastic function. *Journal of Economic Dynamics and Control*, 26, 1379-1396.

BREUSCH, T.S. & PAGAN, A.R. (1979). A simple test for heteroskedasticity and random coefficient variation. *Econometrica*, 47, 1287-1294.

CRIBARI-NETO, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics and Data Analysis*, 45, 215-233.

CRIBARI-NETO, F., FERRARI, S.L.P. & CORDEIRO, G.M. (2000). Improved heterocedasticity-consistent covariance matrix estimators. *Biometrika*, 87, 907-918.

CRIBARI-NETO, F. & ZARKOS, S.G. (1999). Bootstrap methods for heteroskedastic regression models: evidence on estimation and testing. *Econometric Reviews*, 18, 211-28.

CRIBARI-NETO, F. & ZARKOS, S.G. (2001). Heteroskedasticity-consistent covariance matrix estimation: White's estimator and the bootstrap. *Journal of Statistical Computation and Simulation*, 68, 391-411.

CRIBARI--NETO, F. & NILA GALVÃO. (2003). A Class of Improved Heteroskedasticity-Consistent Covariance Matrix Estimators. *Communications in Statistics, Theory and Methods*. 32, 1951-1980.

DAVIDSON, R. & MACKINNON, J.G. (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.

DOORNIK, J.A. (2001). *Ox: an Object--oriented Matrix Programming Language*, 4 ed. Londres: Timberlake Consultants e Oxford: www.nuff.ox.ac.uk/Users/Doornik.

EICKER, F. (1963). Assymtotic normality and consistency of the least squares estimators for families of linear regressions. *Annals of Mathematical Statistics*, 34, 447-456.

GREENE, W.H. (1997). Econometric Analysis, 3 ed. Upper Saddle River: Prentice Hall.

HINKLEY, D.V. (1977). Jackknifing in unbalanced situations. *Technometrics*, 19, 285-292.

HORN, S.D., HORN, R.A. & DUNCAN, D.B. (1975). Estimating heteroskedastic variances in linear models. *Journal of the American Statistical Association*, 70, 380-385.

JUDGE, G.C, HILL, R.C., GRIFFITHS, W.E., LUTKEPOHL, H. & LEE, T.C. (1988). *Introduction to the Theory and Pratice of Econometrics*, 2 ed. New York: Wiley.

LONG, J.S. & ERVIN, L.H. (2000). Using heteroskedasticity-consistent standard errors in the linear regression model. *The American Statistician*, 54, 217-224.

MACKINNON, J.G. & WHITE, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite-sample properties. *Journal of Econometrics*, 29, 305-25.

RAO, C.R. (1973). Linear Statistical Inference and its Applications, 2 ed. New York: Wiley.

WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817-38.

WHITE, H. (1984). *Asymptotic Theory for Econometricians*. Orlando: Academic Press.

Agradecimentos

Os autores agradecem a Francisco Cribari-Neto pela motivação, apoio e pelas pertinentes recomendações ao longo deste trabalho. Os autores agradecem também à CAPES e ao CNPq pelo apoio financeiro.

Abstract

In this paper, we examine the behavior of some well known heteroskedasticity-consistent covariance matrix estimators. We also propose new estimators. Using Monte Carlo simulation, we analize the total relative bias and the total mean square errors of all estimators considered. We also investigate the performance of quasi-*t* tests based on the different estimators in two scenarios, namely: with and without high leverage points in the design matrix.

Keywords: Heteroskesdasticity, quasi-*t* test, high leverage points.

Alguns aspectos da modelagem de dados espacialmente referenciados

Alexandra M. Schmidt*
Aline A. Nobre*
Gustavo S. Ferreira*

Resumo

O avanço de técnicas de simulação estocástica e, paralelamente, de ferramentas computacionais vêm permitindo o uso de modelos altamente estruturados nas diversas áreas da Estatística. Em particular, a área de estatística espacial vem recebendo grande atenção por parte de renomados pesquisadores e modelos mais realistas têm sido propostos.

Este artigo tem como objetivo resumir algumas técnicas para a modelagem de dados espacialmente referenciados. Em particular, revisitamos os modelos para observações feitas em alguns pontos de uma região de interesse e, também, os modelos auto-regressivos espaciais para dados de área, e apresentamos suas estimações seguindo o Paradigma de Bayes. Algumas características dos dois tipos de observações são discutidas e recentes tópicos de pesquisa são apontados. Dois exemplos são analisados, a precipitação da chuva na cidade do Rio de Janeiro, no mês de janeiro de 2002, e o número de casos de malária no Estado do Pará, no mês de janeiro de 1997.

Palavras-chave: Anisotropia; dados de área; geoestatística; modelos auto-regressivos; paradigma de Bayes; processo Gaussiano; e Sistema de Informação Geográfico - SIG.

* Endereço para correspondência: Deptº de Métodos Estatísticos – UFRJ.

1. Introdução

Nos últimos 13 anos, a Inferência Bayesiana vem sofrendo um grande desenvolvimento devido ao avanço de técnicas de simulação estocástica e, também, devido à disponibilização de computadores pessoais velozes e relativamente baratos. Desde a publicação do artigo Gelfand e Smith (1990) vemos em diversas áreas da Estatística propostas de modelos altamente estruturados, que procuram refletir melhor a realidade dos processos sob estudo. Em particular, tem havido um grande desenvolvimento de novas técnicas para o estudo de dados observados ao longo de uma região geográfica, como, por exemplo, o número de casos de doenças respiratórias numa cidade, a modelagem de poluentes do ar num grande centro urbano, etc. Também do ponto de vista geográfico, podemos observar o avanço de diversas técnicas. Os Sistemas de Informação Geográficos - SIGs têm facilitado a visualização e armazenamento de informações relevantes. Um SIG é um sistema automático que codifica, gerencia e analisa dados espaciais. Um SIG pode ser pensado como um mapeamento computadorizado e análise de sistema que usa as características espaciais dos dados para determinar importantes relações como distância, proximidade, etc. Um componente importante de um SIG é um software que ajuda a lidar com dados espacialmente referenciados e visualização de mapas. Um exemplo de software que tem sido muito utilizado é o **Arcview** e uma referência para aprender a usá-lo é Breslin et al. (1999). Do ponto de vista estatístico, o interesse é fazer inferência sobre as informações provenientes desses sistemas e não fazer uma análise puramente descritiva dessas observações.

Cressie (1993) divide a estatística espacial de acordo com os tipos de observações associadas ao espaço em que elas são observadas. De uma forma geral, a estatística espacial contém três grandes áreas: geoestatística, dados de área e processos pontuais. Aqui discutiremos alguns aspectos dos dois primeiros tipos. Para uma discussão detalhada sobre processos pontuais o leitor interessado deve dirigir-se a Diggle (2002). Este artigo está organizado da seguinte forma: a seção seguinte discute princípios básicos de geoestatística, introduzindo processos estacionários (a distribuição espacial é invariável quando a origem do sistema de coordenadas onde as estações se encontram é transladada e isotrópicos, e o processo é estacionário sob rotações em torno da origem). Sob o enfoque Bayesiano, o procedimento de interpolação espacial é descrito. Alguns recentes modelos para tratar de dados espacialmente não-estacionários são mencionados na subseção 2.2. A seção 3 descreve alguns modelos para dados associados a áreas contidas numa região de interesse. Aí é feito um paralelo

com modelos de séries temporais e modelos auto-regressivos simultâneos e condicionais que são discutidos. O procedimento de inferência, sob o enfoque Bayesiano, para tais observações também é descrito. A seção 4 analisa dados de chuva na cidade do Rio de Janeiro e casos de malária no Estado do Pará. Finalmente, a seção 5 levanta alguns outros pontos recentes de interesse de pesquisa na área.

2. Geoestatística

Problemas de geoestatística podem ser encontrados em diferentes áreas da ciência, tais como: meio ambiente, mercado imobiliário, geologia, processamento de imagens, etc. Em geral assume-se que $\{y(x) : x \in G\}$ é uma realização parcial do processo aleatório (ou *campo aleatório*)

$$\{Y(x) : x \in G\},$$

onde G é um subconjunto fixo do $\Re^p$ com volume p-dimensional positivo (Cressie, 1993). Usualmente $p = 1$, 2 ou 3. Em outras palavras, x varia continuamente ao longo da região G. Assuma que $\mu(x) = E(Y(x))$ existe para todo $x \in G$. Usualmente, diz-se que $\mu(.)$ é a tendência do processo espacial $Y(.)$. Suponha também que a variância de $Y(x)$, $V(Y(x))$, existe para todo $x \in G$. Por definição, o processo espacial $Y(.)$ é *intrinsecamente estacionário* se

$$\begin{aligned} &E(Y(x+h) - Y(x)) = 0 \quad e \\ &V(Y(x+h) - Y(x)) = 2\gamma(h), \quad \forall x, x+h \in G, \end{aligned} \tag{1}$$

onde a quantidade $\gamma(.)$ é uma função condicionalmente negativa definida. O processo $Y(.)$ é dito ser *estacionário de segunda ordem* (ou *fracamente estacionário*) se $\mu(x) = \mu$, $\forall x \in G$, isto é, $\mu(x)$ é constante para todo $x \in G$ e a covariância entre dois pontos quaisquer em G é função apenas da diferença entre as duas localizações, isto é,

$$cov\{Y(x), Y(x')\} = c(x - x') \quad \forall x, x' \in G \tag{2}$$

A quantidade $2\gamma(.)$, em (1), é conhecida como variograma e é um dos parâmetros mais importantes na modelagem de geoestatística (Cressie, 1993). A função $c(.)$ é chamada de covariograma. Se em (1), $2\gamma(x - x')$ depender apenas da distância euclideana entre x e x',

$\|x - x'\|$, o processo $Y(.)$ é então chamado de *isotrópico*. Um processo intrinsecamente estacionário e *isotrópico* é chamado de *homogêneo* (Smith, 1996). Se uma dessas condições não se aplica, o processo é *heterogêneo*.

Em geral, assume-se que a variável aleatória $Y(.)$ segue um processo Gaussiano. Mais especificamente, a quantidade aleatória $Y(.)$, que assume valores $y(x)$ para $x \in G$, segue um processo Gaussiano (PG) com média $\mu(.)$ e função de covariância $c(.,.)$, denotado por

$$Y(.) \sim PG\ (\mu(.), c(.,.)),$$

se para quaisquer $x_1, \ldots, x_2 \in G$, e qualquer $n = 1,2\ldots$, a distribuição conjunta de $Y(x_1), \ldots, Y(x_n)$ é uma normal multivariada com parâmetros dados por $E(Y(x_i)) = \mu(x_i)$ e $cov\{Y(x_i), Y(x_j)\} = c(x_i, x_j)$ (O' Hagan, 1994).

Quando um processo é homogêneo, sua variância é constante ao longo de G, isto é, $V(Y(x)) = \sigma^2 \ \forall \ x \in G$. Portanto, podemos escrever a função de covariância de $Y(.)$ como $c(x,x') = \sigma^2 \rho(||x-x'||;\phi^*)$, onde $\rho(.;\phi^*)$ denota uma função de correlação *válida* (positiva definida) em $\Re^p$ que depende possivelmente de um vetor paramétrico ϕ^*. Note que, neste caso, o variograma pode ser escrito como $2\gamma(h) = Var(Y(x)-Y(x')) = \sigma^2 + \sigma^2 - 2\sigma^2\rho(||x-x'||;\phi^*) = 2\sigma^2(1-\rho(||x-x'||;\phi^*))$. Dessa forma, a conveniência dos processos homogêneos fica clara, já que a estrutura de covariância do processo $Y(.)$ pode ser modelada apenas através dos parâmetros σ^2 e ϕ^*. Além disso, modelando $Y(.)$ através de um processo Gaussiano, precisamos especificar apenas seu primeiro e segundo momentos.

Existem na literatura algumas famílias de funções de correlação. Uma função de correlação que é válida em $\Re^{p1}$ também é válida em $\Re^{p2}$ para $p_2 < p_1$. Entretanto, a recíproca não é necessariamente verdadeira (Cressie, 1993). Diggle e Ribeiro Jr (2000) apresentam alguns dos principais modelos paramétricos de funções de correlação assim como simulações de processos Gaussianos univariados mostrando o efeito do uso das diferentes famílias. Dois dos principais exemplos de funções de correlação são:

(a) família exponencial potência:

$$\rho(d;\phi) = exp\left(-(\phi\, d)^{\kappa}\right), \tag{3}$$

onde $\phi^* = (\phi, \kappa)$ com $\phi > 0$ e $\kappa \in (0,2]$. O parâmetro ϕ é de escala e d é a distância euclideana entre dois pontos quaisquer em G. Quando $\kappa = 1$ temos o caso particular da função de correlação exponencial e $\kappa = 2$ corresponde à função de correlação Gaussiana;

(b) família Matérn:

$$\rho(d;\phi,\lambda) = \frac{1}{2^{\lambda-1}\Gamma(\lambda)}\left(\frac{2\sqrt{\lambda}\,d}{\phi}\right)\kappa_\lambda\left(\frac{2\sqrt{\lambda}\,d}{\phi}\right),$$

onde $\phi^* = (\phi,\lambda)$, $\phi > 0$ é o parâmetro de escala, $\lambda > 0$ é o parâmetro de forma. A função $\Gamma(.)$ é a função Gama usual e κ_λ é a função modificada de Bessel do terceiro tipo de ordem λ.

Um aspecto importante de superfícies espaciais é o seu grau de suavidade. Matematicamente essa propriedade é descrita através do grau de diferenciabilidade do processo. A especificação da família de função de correlação é de grande importância, pois, em processos Gaussianos, a suavidade do processo está diretamente relacionada à diferenciabilidade da sua estrutura de covariância. Por exemplo, a função de correlação Gaussiana resulta em processos que são infinitamente diferenciáveis, em outras palavras, em processos extremamente suaves, o que na prática é difícil de se observar. Recentemente, a função de correlação em (b) tem sido a mais usada na literatura, pois além das funções exponencial (quando $\lambda = 0.5$) e Gaussiana (quando $\lambda \to \infty$) serem seus casos particulares, o parâmetro λ controla o grau de diferenciabilidade do processo.

Na subseção a seguir, discutiremos o ajuste de processos Gaussianos sob o enfoque Bayesiano.

2.1 Interpolação espacial bayesiana

Em geral, o principal objetivo em geoestatística é a previsão do processo em localizações não observadas da região G. Em outras palavras, baseado na informação observada de $Y(.)$ em n pontos $(x_1, \ldots, x_n)$ de G, deseja-se prever o processo em k pontos não medidos. Na literatura de geoestatística esse procedimento é conhecido como *krigagem*. Esse termo foi introduzido por Mathéron em diversas publicações, como uma homenagem ao engenheiro sul-africano D.G. Krige que na década de 1950 desenvolveu métodos empíricos para determinar a distribuição de minérios, baseado na informação observada das concentrações de minério num conjunto finito de pontos observados (amostrados). Em geral, krigagem depende das propriedades do momento de segunda ordem de $Y(.)$. Esta subseção destina-se a discutir o procedimento Bayesiano de

inferência dos parâmetros de interesse nos modelos descritos anteriormente. Algumas referências clássicas que descrevem procedimentos de estimação baseados no método da máxima verossimilhança, método dos momentos, etc. são: Journel e Huijbregts (1978), Ripley (1981), Cressie (1993), Stein (1999), para citar alguns.

Assumindo que o processo $Y(.)$ pode ser modelado através de (2), o objetivo então é estimar as quantidades desconhecidas envolvidas em $\mu(.)$ e $c(.,.)$. Do ponto de vista Bayesiano, o modelo é especificado através de, pelo menos, dois níveis de hierarquia. No primeiro nível, encontra-se a distribuição geradora das observações, que depende de quantidades desconhecidas (parâmetros). Sob o enfoque Bayesiano esses parâmetros são quantidades aleatórias, portanto, no segundo nível, especifica-se uma distribuição dos parâmetros antes de se observar os dados (distribuição *a priori*). Esta distribuição depende possivelmente de outros parâmetros, que como não são os principais do modelo, são chamados de hiperparâmetros. As distribuições *a priori* dos hiperparâmetros compreendem os níveis mais baixos de hierarquia do modelo. O objetivo final é obter uma atualização da distribuição desses parâmetros à luz dos dados observados (distribuição *a posteriori*). Uma vez que essa atualização é feita, previsões são feitas com base na distribuição *a posteriori* dos parâmetros.

Omre et al. (1989) discutem uma aproximação Bayesiana para krigagem. Basicamente, eles apresentam dois modelos nos quais eles definem a esperança e a covariância de uma função aleatória. Os parâmetros da função de covariância são assumidos conhecidos e eles fazem uso do método Bayesiano empírico (Migon e Gamerman, 1999) para fixar esses valores. Handcock e Stein (1993) apresentam uma análise Bayesiana completa de krigagem. Assumindo que o processo $Y(.)$ é Gaussiano, eles obtém a distribuição preditiva para uma localidade não observada levando em conta a incerteza associada aos parâmetros da função de covariância. Ribeiro Jr. e Diggle (2002) adotam um modelo espacial linear hierárquico, cujas componentes são processos Gaussianos com uma determinada estrutura paramétrica de covariância. Eles fornecem um catálogo de *posterioris* e distribuições preditivas para particulares combinações de escolha das *prioris*.

Usualmente, no contexto de geoestatística, dadas as observações do processo de interesse em n localizações, $Y = (Y(x_1),\ldots,Y(x_n))$, assume-se que

$$Y \mid \mu, \Sigma \sim N_n(\mu, \Sigma), \tag{4}$$

onde N_n representa a distribuição normal multivariada de dimensão n, μ é um vetor de dimensão n representando a média do processo, e Σ é uma matriz $n \times n$ que representa a estrutura de covariância.

De uma forma geral, podemos supor que a tendência do processo $Y(.)$ não é constante ao longo de G, mas sim uma combinação linear desconhecida de funções conhecidas $f(x)=\{f_1(x),\ldots,f_r(x)\}'$, $x \in G$ isto é, $\mu(x) = \beta' f(x)$, onde $\beta=(\beta_1, \ldots,\beta_r)'$. As quantidades $f_j(.)$, $j=1,\ldots,r$ representam covariáveis que possivelmente explicam o nível de $Y(.)$. Dessa forma esse não é um processo estacionário, já que o nível de $Y(.)$ varia com as localizações em G. Na literatura de geoestatística, a interpolação espacial nesse contexto é conhecida como krigagem universal. Quando a média $\mu(.)$ é assumida constante para todo x em G a interpolação espacial é chamada de krigagem ordinária (Cressie, 1993). No contexto Bayesiano, essa abordagem pode ser descrita assumindo que o processo de interesse $Y(.)$ é descrito pela soma de três componentes, isto é:

$$Y(x) = \beta' f(x) + Z(x) + \varepsilon(x). \tag{5}$$

A primeira componente é descrita por uma tendência polinomial, $\beta' f(x)$, a segunda é um efeito aleatório com estrutura de correlação espacial $Z(x)$, e a última é um ruído branco, $\varepsilon(x)$, com distribuição normal, média 0 e variância constante τ^2, isto é, $\varepsilon(x) \sim N(0, \tau^2)$. Para $Z(.)$ assume-se *a priori* que:

$$Z(x)|\sigma 2,\ \phi^* \sim PG(0,\ \sigma 2\rho(||x-x'||;\ \phi^*)), \tag{6}$$

onde σ^2 é a variância do processo $Z(.)$, $\rho(.;\ \phi^*)$ representa uma função de correlação que depende do parâmetro ϕ^* e $||.||$ denota distância Euclideana. Note que a componente espacial não entra diretamente no processo $Y(.)$, mas sim na sua estrutura de primeira ordem. A componente de ruído, $\varepsilon(.)$, descreve um erro de medida, em outras palavras, caso sejam feitas diversas observações, o resultado vai flutuar em torno do verdadeiro valor (Cressie, 1993). Na literatura de geoestatística, a variância dessa componente é conhecida como *efeito pepita*. Dessa forma, obtemos que a variância de $Y(x)$ condicional a β, σ^2, ϕ^* e τ^2, é dada por

$$V(Y(x)) = \tau^2 + \sigma^2 \quad e \quad Cov(Y(x), Y(x')) = \tau^2 + \sigma^2(1-\rho(||x-x'||;\ \phi^*)),$$

e cada elemento da matriz Σ é dado por $\Sigma_{ij} = \tau^2 + \sigma^2(1-\rho(||x-x'||;\ \phi^*))$. A especificação Bayesiana se completa ao associar as distribuições *a priori* para os parâmetros do vetor $\theta=(\beta,\ \sigma^2,\ \phi^*,\ \tau^2)$.

É razoável assumir que os parâmetros em θ são independentes *a priori*. Como os coeficientes em β representam os efeitos das covariáveis $f(x)$ sobre a média de $Y(.)$, geralmente assume-se que $\beta_i \sim N(0, \sigma^2_\beta)$, onde σ^2_β é uma quantidade fixa conhecida. Quanto maior o valor de σ^2_β mais vaga a informação *a priori* sobre β_i, $i=1,....,r$. Para os parâmetros σ^2 e τ^2 geralmente associa-se uma *priori* gama com uma determinada média e variância. Os parâmetros em ϕ^* representam aqueles envolvidos na função de correlação especificada para o modelo. Por exemplo, no caso da função de correlação exponencial, em (3), uma *priori* usualmente atribuída a ϕ é uma distribuição gama com uma variância grande. A escolha da média dessa gama pode ser difícil. Uma sugestão é atribuir uma média tal que quando a correlação é *0,05*, ϕ é igual a metade da distância máxima *(0,5dmax)* presente na amostra. Em outras palavras, essa *priori* reflete o fato de esperarmos que para distâncias maiores que *0,5dmax* a correlação espacial é próxima de *0*.

Seguindo o paradigma de Bayes, sabemos que a distribuição *a posteriori* de θ, $\pi(\theta|x)$, é proporcional ao produto da função de verossimilhança, $f_n(x|\theta)$, pela *priori*, $\pi(\theta)$, isto é,

$$\begin{aligned} \pi(\theta \mid x) &\propto f_n(x \mid \theta)\pi(\theta) \\ &\propto |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(y - X\beta)' \Sigma^{-1}(y - X\beta)\right\}\pi(\beta)\pi(\tau^2)\pi(\sigma^2)\pi(\phi^*) \end{aligned} \qquad \ldots\ldots \quad (7)$$

Quaisquer que sejam as distribuições *a priori* associadas para os elementos de θ, não é possível fazer nenhuma sumarização da distribuição *a posteriori* acima de forma analítica. Portanto, é preciso fazer uso de métodos de simulação estocástica para se obter amostras da densidade em (7). Na última década, a inferência Bayesiana vem experimentando um grande avanço devido à introdução de métodos de Monte Carlo via Cadeias de Markov (MCMC) e também devido à disponibilidade de computadores velozes. O MCMC é uma técnica poderosa que vem permitindo a análise de modelos altamente estruturados. Nesse contexto, os métodos de simulação estocástica mais utilizados são o amostrador de Gibbs e o Metropolis-Hastings. Para maiores detalhes veja Gamerman (1997). Em termos de implementação desses modelos há uma preocupação com relação à dimensão da matriz Σ. Quanto maior o valor de n maior a dimensão de Σ. Qualquer método de aproximação utilizado para se obter amostras da distribuição em (7) vai envolver o cálculo do determinante e da inversa de Σ que podem demandar muito tempo computacional.

Paralelamente ao desenvolvimento de técnicas de simulação estocástica, surgiram também programas de computadores que geram amostras da *posteriori* de modelos altamente

estruturados. O projeto WinBugs[1] (Spiegelhalter et al., 2002) vem permitindo a implementação de modelos hierárquicos relativamente complexos e, particularmente, de modelos espaciais. Um outro exemplo de programa para obtenção da *posteriori* de modelos geoestatísticos espaciais é o geo-R[2] (Ribeiro Jr. e Diggle, 2001).

2.1.1 Previsão de um vetor de localizações não medidas

Em geoestatística, o maior interesse encontra-se na previsão do processo em pontos não observados. De acordo com o modelo em (4), as observações $Y(x_i)$ estão sendo geradas de acordo com $N_n(\mu, \Sigma)$. Suponha que estamos interessados em prever o processo espacial em um vetor de localizações não medidas $Y_u = (Y(x_{u1}), \ldots, Y(x_{uk}))'$ em pontos $x_{u1}, \ldots, x_{uk}$. O objetivo é obter a distribuição preditiva de $(Y_u \mid Y)$, que é dada por:

$$\ldots\ldots \qquad p(Y_u \mid Y) = \int_\theta p(Y_u \mid Y, \theta)\pi(\theta \mid Y)d\theta. \qquad (8)$$

Seguindo os resultados da teoria normal multivariada (Anderson, 1984), segue que a distribuição conjunta de Y e Y_u é dada por

$$\ldots\ldots \qquad \begin{pmatrix} Y_u \\ Y \end{pmatrix} \mid \theta \sim N_{n+k}\left(\begin{pmatrix} \mu_u \\ \mu \end{pmatrix}; \begin{pmatrix} \Sigma_u & \Psi' \\ \Psi & \Sigma \end{pmatrix}\right), \qquad (9)$$

onde μ_u é um vetor de dimensão k com as médias das respectivas localizações não medidas; μ é um vetor contendo as médias dos n pontos observados; Σ_u é uma matriz de dimensão k e cada elemento representa a covariância entre os pontos não medidos. Cada linha da matriz Ψ, $n \times k$, representa a covariância entre a i-ésima localização medida e a j-ésima não medida, $i = 1, \ldots, n$ e $j = 1, \ldots, k$. Como antes, Σ é a matriz de covariância $n \times n$ das localizações observadas. Da teoria da distribuição normal multivariada, temos que

$$\ldots \qquad (Y_u \mid Y, \theta) \sim N_k\left(\mu_u + \Psi'\Sigma^{-1}(Y - \mu); \Sigma_u - \Psi'\Sigma^{-1}\Psi\right) \qquad (10)$$

---

[1] O WinBugs pode ser obtido gratuitamente da página : http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml.

[2] O geo-R pode ser obtido gratuitamente da página http://www.est.ufpr.br/geoR.

A integração em (8) não tem solução analítica, entretanto aproximações podem ser facilmente obtidas usando métodos de Monte Carlo (Gamerman,1997). Para cada amostra l, $l=1,...,L$, obtida no algoritmo de MCMC, podemos obter uma aproximação para (8), amostrando da distribuição em (10) e calculando

$$p(Y_u \mid Y) \approx \frac{1}{L}\sum_{l=1}^{L} p\left(Y_u \mid \theta^l\right) \qquad (11)$$

Podemos usar a média amostral de Y_u para se obter uma estimativa da esperança de $Y_u \mid Y$, a variância dessa estimativa também pode ser facilmente calculada. Além disso, regiões preditivas de $100(1-\alpha)\%$ podem ser obtidas para $Y_u \mid Y$. Particularmente, regiões de máxima densidade *a posteriori* C, podem ser construídas de modo que

$$C = \{Y_u : p(Y_u \mid Y \geq s(\alpha))\},$$

onde $s(\alpha)$ é a maior constante que garante que $P(Y_u \in C \mid Y) \geq 1-\alpha$.

2.2 Alguns recentes modelos para processos espaciais heterogêneos

Em muitos problemas de geoestatística não é realista assumir homogeneidade do campo aleatório sob estudo. Em outras palavras, não é razoável assumir que a distribuição espacial é invariável quando a origem do sistema de coordenadas onde as estações se encontram é transladada, e que o processo é estacionário sob rotações em torno da origem. Por exemplo, em problemas ambientais, há pouca razão para assumir que tais processos sejam homogêneos, porque acredita-se que há influências locais na estrutura de correlação do processo espacial aleatório (como, por exemplo, efeitos topográficos, impactos de poluentes sobre pequenas vizinhanças das estações monitoradoras, etc.). Na última década tem havido um grande esforço por parte dos pesquisadores dessa área em propor modelos que não assumam homogeneidade do processo espacial sob estudo. Esta seção discutirá algumas das principais referências surgidas nos últimos dez anos nessa área.

Durante os anos de 1980 e 1990, Paul Sampson e Peter Guttorp, da Universidade de Washington, foram pioneiros ao propor uma aproximação para o problema de heterogeneidade espacial. Esta proposta recebeu grande atenção da comunidade estatística depois da publicação do artigo Sampson e Guttorp (1992). A idéia deles está baseada na estimação de um

espaço latente, chamado por eles de espaço D, que denota dispersão. A idéia principal está baseada numa transformação não-linear do espaço amostral, que eles denotam de espaço G (geográfico), para um espaço D, no qual a estrutura espacial é estacionária e isotrópica. Para se obter as localizações dos pontos observados no espaço D, Sampson e Guttorp fazem uso de escalonamento multidimensional (MDS) (Anderson, 1984). Em outras palavras, eles estimam as localizações medidas no espaço D de modo que as correlações observadas se ajustem a distância euclideana entre os pontos em D. Uma vez que as localizações medidas são obtidas no espaço D, o passo seguinte é estimar a realização do processo espacial numa localização não medida, e Sampson e Guttorp fazem uso de *thin-plate splines* (Green e Silverman, 1994) para atingir esse objetivo. Entretanto, há algumas críticas a essa proposta, pois uma vez que os pontos são obtidos no espaço D, eles são fixados e a interpolação para localizações não medidas é baseada nessas localizações fixas sem levar em conta a incerteza associada ao mapeamento.

Le e Zidek (1992) propõem uma alternativa Bayesiana à krigagem na qual a incerteza sobre a matriz de covariâncias Σ é naturalmente refletida. Eles definem uma distribuição *a priori* Wishart invertida para Σ com matriz de escala Ψ. Note que dessa forma a matriz Σ, não necessariamente, tem uma forma isotrópica já que sua parametrização não envolve a distância euclideana entre os pontos em G. A validade da matriz de covariâncias é garantida devido ao uso da Wishart invertida como *priori* de Σ. Uma sugestão para fixar a matriz de escala Ψ é usar o método não-paramétrico de Sampson e Guttorp para obter uma estimativa para Ψ. Um exemplo dessa idéia pode ser visto em Sun et al. (1998).

Ainda no contexto da deformação espacial, Schmidt e O'Hagan (2003) propõem uma aproximação Bayesiana para o modelo inicialmente proposto por Sampson & Guttorp. O mapeamento dos pontos observados para o espaço latente D se dá através do uso de um processo Gaussiano como distribuição *a priori* para a função $d(.)$, que leva as localizações do espaço geográfico G para o espaço latente D. Uma vez que a distribuição *a posteriori* é obtida, previsões para localizações não medidas podem ser obtidas levando em conta toda a incerteza associada. Neste contexto, toda a inferência é feita numa única estrutura. Damian et al. (2001) propuseram independentemente de Schmidt e O'Hagan (2003) uma versão semiparamétrica para o modelo de deformação espacial. Schmidt (2001) faz uma comparação detalhada entre os dois métodos, além de descrever outros para a deformação espacial que são baseados no estimador de máxima verossimilhança.

Nos últimos cinco anos, surgiram outras propostas para a modelagem de observações espaciais heterogêneas. Esses modelos são baseados na convolução de processos estacionários. Mais especificamente, Higdon et al., (1999) propõem um modelo que é baseado na especificação

de médias móveis de processos Gaussianos. Isto é, o processo não estacionário é representado por

$$Y(x) = \int_{\Re^2} k_x(u)\varepsilon(u)du,$$

onde $k_x(.)$ denota um kernel que é centrado no ponto x e cuja forma depende da localização x. A forma do kernel é assumida normal, centrada na localização espacial x com matriz de covariância, $\Sigma(x)$, que varia de acordo com a localização, e cujos parâmetros são permitidos a variarem suavemente ao longo da região. A modelagem do processo $Y(.)$ como acima resulta na seguinte estrutura de covariância

$$C(x, x') = \int_{\Re^2} k_x(u)k_{x'}(u)du,$$

que claramente varia com a localização e, portanto, não é homogênea. Além disso, eles fazem uso de modelos Bayesianos hierárquicos de modo que a incerteza de todas as quantidades envolvidas no modelo é considerada.

Por outro lado, Fuentes e Smith (2000) propõem a modelagem de $Y(.)$ através da convolução de processos Gaussianos localmente estacionários, isto é

$$Y(x) = \int_G K(x - s)Z_{\theta(s)}(x)ds,$$

onde $K(.)$ é uma função kernel com *bandwidth* h. O processo $Z_{\theta(x)}$ compreende uma família de processos Gaussianos estacionários indexados por θ. Dessa forma, é simples demonstrar que a covariância do processo $Y(.)$ é dada pela convolução de covariâncias localmente estacionárias, $C_{\theta(s)}(x_1-x_2)$, onde x_1 e x_2 são dois pontos em G, isto é,

$$C(x_1, x_2; \theta) = \int K(x_1 - s)K(x_2 - s)C_{\theta(s)}(x_1 - x_2)ds.$$

Mais recentemente, Kim et al. (2002) propuseram um modelo que lida com mudanças bruscas na estrutura de covariância. Eles propõem uma partição da região sob estudo usando a Tesselagem de Voronoi, de modo que o processo é assumido estacionário dentro de cada sub-região e independente entre as diferentes sub-regiões. Usando o enfoque Bayesiano eles automaticamente obtém uma estimativa da incerteza das fronteiras das sub-regiões, assim como

sobre a forma dessas sub-regiões. Dessa forma, as previsões são baseadas numa média dos diferentes modelos para as diversas sub-regiões, que tendem a suavizar os modelos individuais pouco suaves.

3. Dados de área

Um outro tipo de observação no contexto espacial são as observações de área. Neste caso, as localizações podem ser pontos ou regiões, entretanto na maioria dos casos em que os dados estão relacionados a pontos, usa-se geoestatística. Observações são obtidas a partir de um número finito de localizações que compreendem toda a região sob estudo. Exemplos típicos relacionados aos dados de área são a presença de espécies de uma planta num quadrado, o número de casos de dengue nos diversos bairros da cidade do Rio de Janeiro, etc. Note que a região sob estudo é dividida em sub-regiões regulares (divisão da região numa grade) ou irregulares (bairros, municípios, etc.). Em dados desse tipo, as sub-regiões formam tipicamente uma partição da área de estudo e a possibilidade de uma resposta ocorrer entre localizações é excluída. Neste caso, a idéia é usar modelos que especifiquem que o processo de interesse é influenciado, de alguma forma, pela resposta do mesmo em localizações vizinhas. Os modelos mais populares para dados de área são similares a modelos comumente usados para séries temporais discretas.

Supondo que a área de interesse possa ser dividida em n sub-regiões, regulares ou não, podemos usar a idéia de modelos auto-regressivos temporais e supor que a resposta para cada área i, Z_i, $i=1,\ldots,n$, é uma auto-regressão de primeira ordem na média da resposta dos seus vizinhos, isto é,

$$Z_i = \rho\left(\frac{\sum_{j\in N_i} Z_j}{|N_i|}\right) + \varepsilon_i, \qquad \varepsilon_i \sim N(0,\sigma^2),$$

onde N_i é o conjunto dos vizinhos da área i e $|N_i|$ é o número de vizinhos à área i. Mais geralmente, podemos assumir que existe uma tendência μ no processo que em notação matricial pode ser representado por

$$Z-\mu = \rho W(Z-\mu)+\varepsilon, \qquad \varepsilon \sim N(0,\sigma^2 I_n) \qquad \ldots\ldots \qquad (12)$$

onde W é uma matriz de pesos representando a estrutura de vizinhança e I_n representa a matriz identidade de ordem n. A matriz de vizinhança W pode ser especificada apenas através das localizações adjacentes, isto é, $W_{ij}=1$ se as áreas i e j são adjacentes ($i \sim j$) e $W_{ij}=0$, caso contrário. Também pode-se especificar uma matriz W que tenha pesos diferentes de zero, informando que a resposta na sub-região i não depende apenas daquelas localizações adjacentes. O modelo em (12) é conhecido na literatura como *modelo espacial auto-regressivo*. Assim como no caso de séries temporais, podemos especificar dois tipos de modelos espaciais auto-regressivos. Na sua forma mais geral, eles são como seguem:

1. Modelo Auto-regressivo Simultâneo (SAR)

$$Z_i - \mu_i = \sum_j S_{ij}(Z_j - \mu_j) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2),$$

onde $i=1,\ldots,n$, e $S \equiv \{S_{ij}\}$ é tal que I_n-S é não-singular. Na forma matricial podemos escrever

$$Z - \mu = S(Z - \mu) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n)$$

2. Modelo Auto-regressivo Condicional (CAR)

Neste caso, especifica-se a distribuição condicional do processo na área i dados os vizinhos, isto é

$$\ldots\ldots \qquad (Z_i \mid Z_j, j \neq i) \sim N(\mu_i + \sum_j C_{ij}(Z_j - \mu_j), \sigma^2), \qquad (13)$$

onde $C \equiv \{C_{ij}\}$ é tal que I_n-C é simétrica e positiva definida. Equivalentemente,

$$Z_i - \mu_i = \sum_j C_{ij}(Z_j - \mu_j) + \nu_i,$$

onde $\nu_i \sim N(0, \tau^2)$, para $i=1,\ldots,n$.

Entretanto, em contraste com a abordagem para séries temporais, as duas especificações fornecem dois diferentes modelos, isto é, mesmo se fizermos $C_{ij}=S_{ij}$ o modelo CAR fornece uma distribuição conjunta que é diferente daquela do SAR (Cressie,1993).

Seguindo as especificações acima, a distribuição dos dados é da seguinte forma:

$$Z \sim \begin{cases} N(\mu, (I-S)^{-1}\Lambda(I-S')^{-1}), & \text{para o SAR} \\ N(\mu, (I-C)^{-1}M), & \text{para o CAR,} \end{cases}$$

onde $\Lambda = diag(\sigma^2, \ldots, \sigma^2)$ e $M = diag(\tau_1^2, \ldots, \tau_n^2)$. Esta demonstração pode ser vista em Cressie (1993) pág.: 414. Se em particular, $\tau_i^2 = \tau^2$, $\forall i = 1,2,\ldots,n$, o log da função de verossimilhança associada ao processo vindo de um dos dois modelos é dado por:

$$-\frac{n}{2}log(2\pi\sigma^2) + \frac{1}{2}|B| - \frac{1}{2\sigma^2}(Z-\mu)'B(Z-\mu) \qquad .. \qquad (14)$$

onde

$$\mathbf{B} = \begin{cases} (I-S')(I-S), & \text{para um SAR} \\ (I-C), & \text{para um CAR.} \end{cases}$$

Cressie (1993) apresenta os estimadores de máxima verossimilhança obtidos a partir da função em (14).

Comparando algumas propriedades dos dois modelos podemos decidir qual usar. No SAR os termos ε_i são correlacionados com $\{Z(s_j): j \neq i\}$, mas no CAR v_i não é correlacionado com $\{Z(s_j): j \neq i\}$. Portanto, em termos de estimação e interpretação, o CAR deve ser preferido ao SAR. Além disso, os estimadores de mínimos quadrados do SAR são inconsistentes. Os modelos CAR fornecem o melhor erro quadrático de previsão na interpolação de $Z(s_i)$ baseada nos $\{Z(s_j): j \neq i\}$. Os dois modelos são equivalentes se, e somente se, as suas matrizes de variância-covariância forem iguais, isto é, se

$$(I_n - C)^{-1}M = (I_n - S)^{-1}\Lambda(I_n - S')^{-1}.$$

Como a matriz M é diagonal, qualquer modelo SAR pode ser representado por um modelo CAR, entretanto a recíproca não é necessariamente verdadeira. Além disso, o CAR precisa ter uma matriz de dependência espacial C simétrica, enquanto a matriz S no SAR não é necessariamente simétrica. Isto segue do fato que $(I_n - S')(I_n - S)$ é simétrica. Apesar de isso

parecer uma vantagem do SAR, como os parâmetros em $\{S_{ij}\}$ são estimados através de $(I_n - S')\Lambda(I_n - S)$, isto pode resultar numa possível não-identificabilidade. Uma outra propriedade interessante do CAR é que sua especificação fornece diretamente as distribuições condicionais completas dos parâmetros do modelo, fator determinante para o uso do amostrador de Gibbs em MCMC.

3.1 Inferência bayesiana para modelos CAR

No contexto Bayesiano, geralmente o modelo CAR é usado como informação *a priori* de um parâmetro do modelo para o processo de interesse. Em seguida descreveremos um modelo genérico para exemplificar a modelagem Bayesiana de um processo condicionalmente auto-regressivo. Na subseção 4.2 ilustraremos um exemplo específico desse tipo de abordagem. Como antes, assuma que a área de interesse possa ser dividida em n sub-regiões, regulares ou não, e Z_i representa a quantidade de interesse que observamos em cada sub-região i, $i = 1, \ldots, n$. Um possível modelo para Z_i é:

$$\ldots \qquad Z_i = \mu + \sum_{k=1}^{q} \beta_k X_{ki} + S_i \,, \; i = 1, \ldots, n \qquad (15)$$

onde μ representa um nível geral, comum a toda região sob estudo; $X_i = (X_{1i}, \ldots, X_{qi})$ representa um vetor de possíveis covariáveis que podem explicar o processo e β_k representa o efeito da k-ésima covariável na resposta Z; as componentes $S = (S_1, \ldots, S_n)$ são efeitos aleatórios que podem ser vistos como variáveis latentes que capturam efeitos desconhecidos ou não medidos pelas covariáveis. Caso acreditamos que essas covariáveis não medidas apresentam uma estrutura espacial (por exemplo, efeitos ambientais não medidos) então o modelo para S_i deve permitir a presença de tal estrutura.

Sob o enfoque Bayesiano, a equação (15) representa o primeiro nível de hierarquia do modelo. No segundo nível devemos especificar a distribuição *a priori* do vetor paramétrico $\theta = (\mu, \beta_1, \ldots, \beta_q, S)$. Geralmente assume-se *a priori* que esses parâmetros são independentes e que $\mu, \beta_1, \ldots, \beta_q$ seguem uma distribuição *a priori* normal centrada em zero com variância grande. Dessa forma, deixamos que os dados nos dêem maiores informações sobre tais parâmetros. Já os efeitos aleatórios S_i's assumem uma *priori* auto-regressiva condicional intrínseca (Besag *et al.*, 1991). Essencialmente a estrutura dessa *priori* é como descrita em (13).

Mais especificamente,

$$(S_i \mid S_j = s_j, j \neq i) \sim N(m_i, v_i), \text{ onde} \qquad (16)$$

$$m_i = \frac{\sum_{j \in \delta_i} W_{ij} s_j}{\sum_{j \in \delta_i} W_{ij}} \quad e \quad v_i = \frac{v^*}{\sum_{j \in \delta_i} W_{ij}},$$

onde δ_i representa o conjunto de áreas subjacentes a i. Essa especificação resulta na seguinte distribuição *a priori* conjunta para S:

$$(S \mid v^*) \propto \frac{1}{v^{*n}} exp\left\{ - \frac{1}{2v^{*2}} \sum_{i=1}^{n} \sum_{j<i} W_{ij} (S_i - S_j)^2 \right\},$$

que é uma distribuição imprópria já que é baseada nas diferenças pareadas entre os S_i's, em outras palavras, essa *priori* é invariante à locação. Como *prioris* impróprias podem resultar em posterioris impróprias, na prática impomos uma restrição para que esses efeitos somem zero. A especificação se completa ao determinar a matriz de vizinhança W_{ij} e *a priori* para a variância v^*. É comum assumir que $W_{ij} = 1$ se $i \sim j$ e 0 caso contrário. Nesse caso, temos que

$$m_i = \frac{\sum_{j \in \delta_i} s_j}{n_i} \quad e \quad v_i = \frac{v^*}{n_i},$$

onde n_i é o número de vizinhos da sub-região i. Em outras palavras, a média condicional de S_i, m_i, é dada pela média aritmética dos efeitos dos seus vizinhos; e a variância condicional v_i, é proporcional ao número de vizinhos, daí a denominação CAR intrínseco. Essa especificação é especialmente relevante quando a região é dividida em sub-regiões irregulares. Outras estruturas de vizinhança podem ser adotadas, por exemplo, alguma baseada na distância entre os centróides das sub-regiões. O importante, como mencionado anteriormente, é que esses pesos sejam simétricos. Geralmente assume-se para v^* uma *priori* gama invertida. A inferência pode ser sensível a escolha dessa *priori*.

Uma outra possibilidade nesse tipo de modelagem é considerar na equação (15), além de um efeito espacial S, um efeito independente, U. Esse modelo é conhecido na literatura como de convolução e é descrito detalhadamente em Mollié (1996).

É importante ressaltar que grupos de sub-regiões que formam ilhas e que não têm nenhuma observação causarão problemas devido à impropriedade da *priori*. Para esses efeitos *a posteriori* condicional será proporcional *a priori*, já que a verossimilhança não traz informação sobre essas regiões e, portanto, *a posteriori* será imprópria.

4. Exemplos

Esta seção tem como objetivo *ilustrar* dois exemplos de dados espacialmente referenciados. Inicialmente apresentaremos possíveis propostas de modelagem da precipitação na cidade do Rio de Janeiro para o mês de janeiro de 2002. Baseados na informação sobre a quantidade de chuva proveniente de algumas estações monitoradoras, o objetivo é prever a quantidade de chuva para esse mês, para uma grade de pontos sobreposta sobre o município e assim obter uma superfície que represente a quantidade de chuva.

No exemplo seguinte ilustraremos uma possível modelagem para o número de casos de malária no Estado do Pará, aquele que apresenta o maior número de casos da doença no Brasil. Como os dados são registrados por municípios, faremos uso dos modelos auto-regressivos condicionais para fazer inferência sobre o número de casos de malária e verificar a presença de estrutura espacial no processo sob estudo.

4.1 Análise da precipitação da chuva na cidade do Rio de Janeiro

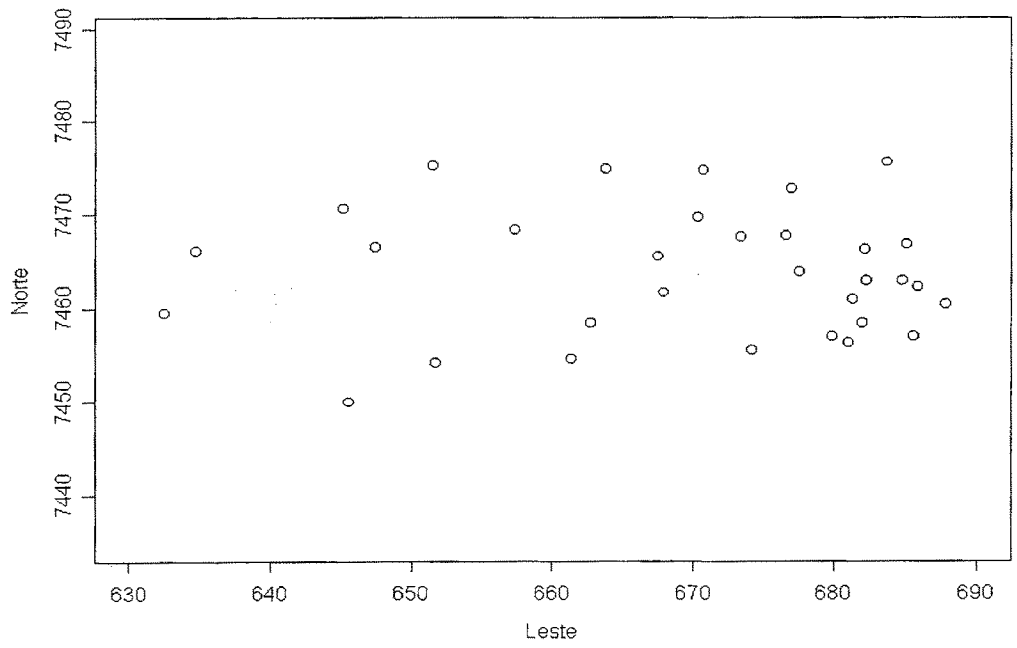
Para esta análise, utilizou-se dados do índice pluviométrico nas quatro primeiras semanas do ano de 2002 na cidade do Rio de Janeiro. Os dados foram fornecidos pela Fundação Geo-Rio (http://www.rio.rj.gov.br/georio/) e contém o índice pluviométrico para *32* estações de monitoramento. A Figura 1 mostra a localização geográfica dessas estações. Utilizaremos a abordagem Bayesiana descrita na subseção 2.1 para prever a quantidade de chuva em locais não observados. Para isto, foi criada uma malha retangular sobre a cidade contendo *868* pontos, e a quantidade de chuva será aí prevista a partir da equação (8). Geralmente, na prática, dados ambientais raramente seguem uma distribuição normal. Assume-se, então, que alguma transformação é feita para obter normalidade. Seguindo a sugestão de Stidd (1973) utilizaremos a raiz cúbica da quantidade de chuva. As únicas covariáveis disponíveis para a explicação da quantidade de chuva são a latitude e longitude das estações monitoradoras. Devido a esse fato, e de acordo com o modelo descrito em (5), três modelos foram inicialmente explorados:

(i) $[chuva(\mathbf{x})]^{1/3} = \beta + Z(\mathbf{x}) + \varepsilon(\mathbf{x})$.

(ii) $[chuva(\mathbf{x})]^{1/3} = \beta + \beta_1 LO(\mathbf{x}) + \beta_2 NS(\mathbf{x}) + Z(\mathbf{x}) + \varepsilon(\mathbf{x})$.

(iii) $[chuva(\mathbf{x})]^{1/3} = \beta + \beta_1 LO(\mathbf{x}) + Z(\mathbf{x}) + \varepsilon(\mathbf{x})$.

Figura 1- Localização das estações monitoradoras de chuva da geo-Rio no Município do Rio de Janeiro.



onde *NS* denota a direção norte-sul e *LO* a direção leste-oeste. Para os três modelos, $Z(.)$ é um processo Gaussiano com média *0* e estrutura de correlação, $\rho(.;\phi,\kappa)$, dada pela função exponencial potência e variância igual a σ^2. A componente $\varepsilon(.)$ representa o erro de medida (efeito pepita) e como antes, assumimos que $\varepsilon(.) \sim N(0,\tau^2)$.

Observou-se que os dados apresentam muita variabilidade na direção leste-oeste, enquanto na direção norte-sul o mesmo não ocorre. O modelo com duas covariáveis, quando utilizado, produz distribuição *a posteriori* de β_2 centrada em zero. Isto nos levou a escolher o modelo (iii) para realizar as predições nos locais não observados. As *prioris* utilizadas foram as seguintes: $\tau^2 \sim Gamma(5,5)$, $1/\sigma^2 \sim Gamma(5,5)$, $\phi \sim Gamma(1.178,1.085)$, $\kappa \sim U(0,2)$, $\beta \sim N(0,10)$, e $\beta_1 \sim N(0,10)$. *A priori* para ϕ é centrada em $3/(0.05dmax)$, onde $dmax = 55.28$.

Para a implementação do modelo utilizamos o programa computacional **WinBugs.** O código utilizado está descrito no apêndice A.1. Utilizou-se uma cadeia de *20 000* iterações onde as *1 000* primeiras foram descartadas e, a partir daí, armazenou-se as observações de *10* em *10* para minimizar problemas de autocorrelação entre as amostras. A convergência das cadeias foi verificada através do uso de duas cadeias que começaram de valores iniciais diferentes. Um estudo rigoroso sobre a convergência das cadeias pode ser feito através do uso do software CODA[3] que possui alguns dos principais métodos para verificação de convergência de amostras obtidas via MCMC.

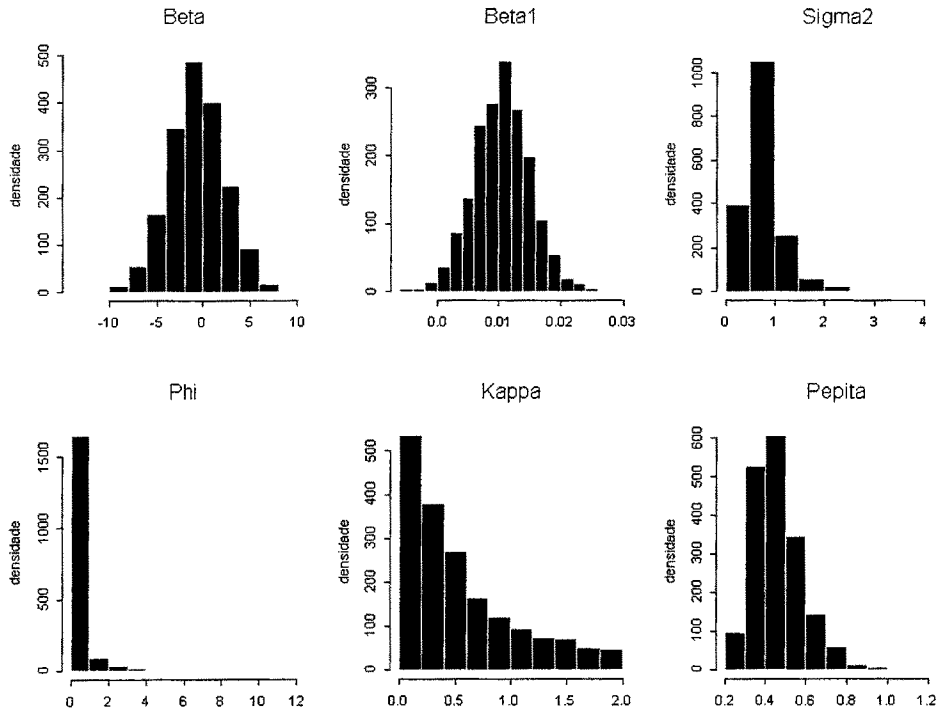
A Tabela 1 mostra o sumário da distribuição *a posteriori* dos parâmetros do modelo (iii), descrevendo claramente a incerteza associada a eles. Observe que o coeficiente da variação no sentido leste-oeste, β_1 é significativo e positivo, além de estar bem concentrado em torno da sua média. A Figura 2 mostra os histogramas das distribuições *a posteriori* dos parâmetros. Note que seguindo a equação (3), verificamos pela distribuição *a posteriori* de ϕ que a componente espacial Z(.) é significativa, já que a mediana *a posteriori* de ϕ é da ordem de 0.12 (pela parametrização em (3) quanto menor o valor de ϕ, maior a correlação espacial). E mais, a variância *a posteriori* dessa componente, σ^2, é maior do que a variância da componente do ruído branco, τ^2, mostrando que ela deve estar presente no modelo.

Tabela1 - Sumário da Distribuição *a Posteriori* dos Parâmetros do Modelo da Chuva

Parâmetro	Média	Mediana	2.5%	97.5%
β	-0.6646	-0.7276	-6.634	5.035
β_1	0.0105	0.0106	0.0011	0.019
$1/\tau^2$	0.4608	0.4404	0.2776	0.755
σ^2	0.7827	0.6760	0.3352	1.907
ϕ	0.3750	0.1197	0.0090	2.309
κ	0.5495	0.3916	0.0110	1.808

[3] O software CODA pode ser obtido da página:
http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml.

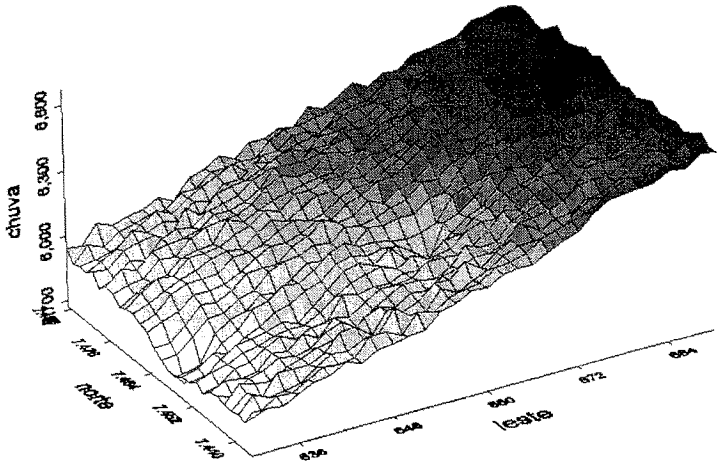
Figura 2 - Distribuição *a posteriori* dos parâmetros do modelo da chuva: $\beta, \beta_1, \sigma^2, \phi, \kappa, \tau^2$ (pepita).



Finalmente, a Figura 3 contém as médias das distribuições preditivas para cada um dos *868* locais de predição. O gráfico gerado fornece o índice pluviométrico com os dados na escala original. Como já era esperado, a quantidade de chuva prevista cresce consideravelmente no sentido leste-oeste.

Note que nos modelos aqui propostos estamos implicitamente assumindo que a chuva no Rio de Janeiro é isotrópica. Entretanto, a geografia da cidade é bem particular, apresentando uma grande cadeia de montanhas, que provavelmente tem influência sobre a estrutura de correlação do processo. Seria interessante investigar a hipótese de isotropia mais a fundo e aplicar um dos modelos descritos na seção 2.1 para esses dados. Outra possibilidade seria utilizar outras covariáveis, tal como altitude, como variável explicativa.

Figura 3 - Média da distribuição preditiva *a posteriori* da chuva para as 868 localidades não medidas da grade, usando o modelo (iii)



4.2 Estimando os casos de malária no Pará

Nesta seção apresentamos um exemplo ilustrativo de dados de área, referente ao número de casos de malária em alguns municípios do Estado do Pará no mês de janeiro de 1997. O Estado do Pará contém aproximadamente 140 municípios, e é o estado que apresenta maior incidência da doença no Brasil. Devido à falta de coleta dos dados, a região de estudo foi reduzida para 69 municípios, dentre os quais 34 não apresentam informação sobre casos de malária (veja Figura 4(a)). Então, o número de casos de malária, y_i, foi modelado como variáveis aleatórias independentes com distribuição de Poisson, isto é,

$$y_i \mid r_i, e_i \sim Poisson(e_i r_i), \quad i = 1,\ldots,69.$$

onde e_i representa a contagem esperada da malária no município i, baseada no tamanho da população e r_i representa o risco relativo da malária no município i. Para considerar a variabiliade dos riscos relativos, assumimos a presença de um efeito aleatório com estrutura espacial, ou seja, estamos assumindo que municípios vizinhos se comportam de maneira similar. Neste caso, o logaritmo dos riscos relativos é modelado da seguinte forma,

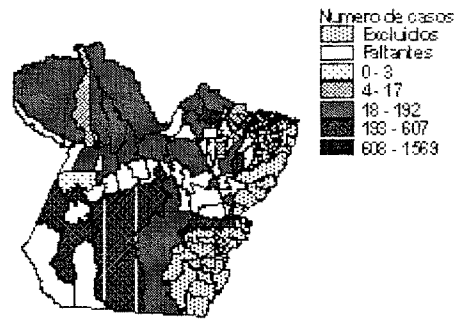
$$log(r_i) = \mu + S_i, \qquad (17)$$

onde μ representa um nível comum a todos os municípios e S_i um efeito aleatório representando a componente espacialmente estruturada com distribuição CAR. A Figura 4(a) apresenta o mapa com o número de casos de malária. Podemos observar que existe uma grande variabilidade no número de casos, alguns municípios apresentam menos de 17 casos, enquanto em outros foram registrados aproximadamente 1500 casos. O risco relativo clássico é dado pela razão entre o número de casos observados e o número de casos esperados em cada município, isto é, $\hat{r}_i = y_i / e_i$. Entretanto, quando a doença é rara e/ou as áreas são pequenas, os mapas de taxas cruas podem ser altamente afetados pela variabilidade, pois o erro padrão estimado de $\hat{r}_i$ é dado por $\sqrt{y_i} / e_i$. O mapa com o risco relativo clássico é mostrado na Figura 4(b). Note que utilizando esta abordagem, não conseguimos fazer qualquer inferência sobre os casos de malária para os 34 municípios faltantes. Com o modelo proposto na seção 3.1, a inferência para os municípios faltantes é alcançada facilmente. Sob o enfoque Bayesiano, a análise é realizada da seguinte forma, assuma o vetor $y = (y_{obs}, y_{mis})$, onde y_{obs} denota os valores observados e y_{mis} os valores faltantes. Uma variável indicadora de mesma dimensão de y é inserida no modelo, tal que $I = 1$ se y é observado e 0 caso contrário. A distribuição dos dados observados, (y_{obs}, I), é obtida integrando em relação à distribuição de y_{mis}. A distribuição dos dados faltantes, é baseada na distribuição *a posteriori* conjunta dos parâmetros do modelo e dos dados faltantes, condicional aos dados observados, resultando em uma amostra de tamanho L, (y^l_{mis}, θ^l), $l = 1,...,L$, e, então, a distribuição *a posteriori* para y_{mis} é obtida integrando com relação aos parâmetros. Para fazer a inferência sobre os dados faltantes no **WinBugs** basta colocar o valor NA para esses municípios sem informação e a distribuição dos y_{mis} é automaticamente obtida. Finalmente, o mapa da Figura 4(c) apresenta a mediana *a posteriori* do risco relativo obtido para o modelo cujo efeito aleatório tem estrutura espacial. Como é de se esperar, o mapa apresenta a noção de vizinhança com o sudoeste do estado, apresentando maior incidência de malária. Observamos claramente o ganho de informação utilizando essa abordagem, já que neste caso conseguimos fazer inferência para todos os 69 municípios.

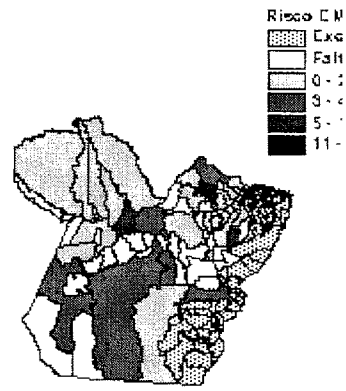
Além de obtermos a previsão dos riscos para os municípios com dados faltantes, sob o enfoque bayesiano, obtemos naturalmente a incerteza associada à estimação de todos os parâmetros do modelo. Com respeito à convergência dos parâmetros para a distribuição objetivo, a saber *a posteriori* de μ e $\boldsymbol{S} = (S_1, S_2, ..., S_n)$, como na seção anterior, utilizamos duas cadeias com valores iniciais diferentes.

Figura 4 - Mapas do Pará representando (a) o número observado de casos de malária, por município, em janeiro de 1997; (b) o risco de malária estimado pelo método da máxima verossimilhança ($\hat{r}_i = y_i / e_i$); (c) mediana *a posteriori* do risco de malária, usando o modelo CAR (exp($\mu + S_i$) na equação (17))

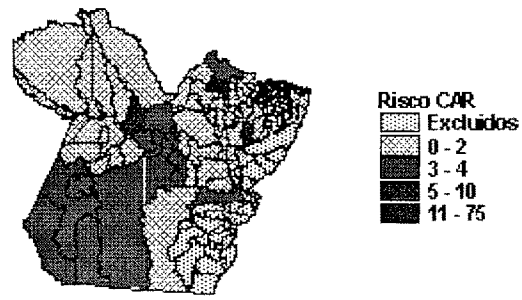
(a)



(b)



(c)



5. Conclusão e outros tópicos de pesquisa

Este artigo teve como objetivo fazer uma breve revisão de processos cuja realização está relacionada a uma localização no espaço. Em particular, tratamos da modelagem de processos aleatórios baseada na observação num conjunto de pontos e num conjunto finito de regiões ou áreas. Alguns aspectos importantes de modelos de geoestatística foram mencionados, assim como o problema da modelagem de dados heterogêneos. Esse tem sido um tópico de muito interesse dos pesquisadores dessa área nos últimos anos.

A modelagem de dados de área tem sido muito utilizada para o mapeamento de doenças, como visto no exemplo da subseção 4.2. Note que uma covariável importante que pode explicar o número de casos de malária é a quantidade de chuva observada em cada área (município) no mesmo período. Assim teríamos um processo pontual servindo como variável explicativa para uma variável resposta que varia por área. A questão então é determinar qual é o valor da quantidade de chuva que representa a quantidade observada em cada município. Esse problema é conhecido na literatura como troca de suporte (Cressie, 1993). Nobre (2003) apresenta uma discussão sobre esse problema no contexto da chuva e malária, e sugere como modelar esses processos conjuntamente.

Existem atualmente diversos tópicos de interesse de pesquisa para dados espaciais. Por exemplo, um aspecto importante é a modelagem de observações não normais. Uma importante referência nessa área é o artigo Diggle et al. (1998) que propõe um modelo para observações provenientes da família exponencial. Já existe disponível uma versão do **geo-R** para dados não normais, o **geo-Rglm** (Christensen e Ribeiro Jr, 2002).

Um outro problema importante é a modelagem de processos espaciais multivariados. Usualmente, em dados ambientais, observa-se mais de um processo no mesmo ponto geográfico. Por exemplo, uma estação monitoradora de poluentes, em geral, terá observações de diferentes poluentes, devido ao custo associado à implantação da monitoração naquele ponto. Portanto, nesse tipo de observação, precisa-se levar em consideração não só a estrutura de covariância espacial, mas, também, a estrutura de covariância entre os diversos processos observados no mesmo ponto. Dois importantes artigos nessa área são Mardia e Goodall (1993) e Brown et al. (1994). Schmidt e Gelfand (2003) propõem um modelo Bayesiano de co-regionalização linear, onde o processo multivariado é modelado a partir da combinação linear de processos espaciais univariados. Finalmente, um outro aspecto que não foi mencionado, até agora, é a consideração da componente temporal para esse tipo de observação. Para os exemplos citados até aqui, é

clara a relação espaço-temporal existente nesses processos. Novamente, considerando os problemas ambientais, geralmente, temos observações feitas em diversos pontos da região de interesse ao longo de diferentes instantes de tempo. O desafio aí é propor uma estrutura de covariância que descreva a correlação espacial entre as observações para cada instante de tempo. Esse é um outro tópico que tem sido de muito interesse dos pesquisadores dessa área. É importante propor uma estrutura de covariância *válida* (matriz positiva-definida). Durante muitos anos explorou-se uma estrutura de covariância que é dada pelo produto da covariância temporal pela covariância espacial. Entretanto, essa hipótese resulta no fato de que a estrutura de covariância resultante é *separável*; isto é, sob essa modelagem, implicitamente assume-se que $cov(Y_t(x), Y_{t'}(x')) = cov(Y_{t'}(x), Y_t(x'))$, o que, no caso de problemas ambientais, não parece refletir realisticamente os processos sob estudo. Cressie and Huang (1999) e Brown et al. (2000) são exemplos de artigos que propõem estruturas de correlação espaço-temporais que são não-separáveis.

A Códigos dos Programas Utilizados no WinBugs

Aqui são apresentados os códigos dos programas utilizados no **WinBugs** para se obter a distribuição *a posteriori* dos parâmetros para o modelo da chuva na cidade do Rio de Janeiro e para o número de casos de malária no Estado do Pará.

A.1 Código do Programa para o Modelo da Precipitação no Rio de Janeiro

```
model
    {
     for (i in 1:N){
     chuva[i]~dnorm(mu[i],pep)
     mu[i]<-beta+beta1*leste[i]+w1[i]
     mu1[i]<-0
     }
     w1[1:N]~spatial.exp(mu1[],leste[1:N],norte[1:N],tau,phi,kappa)
     #prioris:
     pep~dgamma(5,5)
     pepita<-1/pep
     tau~dgamma(5,5)
     sigma2<-1/tau
     #priori para phi centrada em 3/(0.5*dmax)=1.085384 onde
```

```
#dmax=55.28
 phi~dgamma(1.178,1.085384)
 beta~dnorm(0,0.1)
 beta1~dnorm(0,0.1)
 kappa~dunif(0,1.99999)
 #previsão:
 for (j in 1:M){
        mu2[j]<-0
        w2[j]~spatial.unipred(mu2[j],leste2[j],norte2[j],w1[])
        mu3[j]<-w2[j]+beta+beta1*leste2[j]
        chuva.pred[j]~dnorm(mu3[j],pep)
     }
 }
```

A. 2 Código do Programa para o Modelo da Malária no Pará

```
model
{
 b[1:69]~car.normal(adj[], weights[], num[], tau)
 for (i in 1 : 69) {
 Y[i]~dpois(mu[i])
 log(mu[i]) <- log(E[i]) + alpha + b[i]
 RR[i] <- exp(alpha + b[i]) }
#prioris
 alpha~dflat()
 tau~dgamma (0.00625, 0.025)
}
```

Referências bibliográficas

ANDERSON, T. (1984) An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, Inc.

BESAG, J., YORK, J. E MOLLIÉ (1991) Bayesian Image Restoration, with Two Applications on Spatial Statistics. Annals of the Institute of Statistical Mthematics, 43, 1-59.

BRESLIN, P., FRUNZI, N., NAPOLEON, E. E ORMSBY, T. (1999) Getting to know ArcView GIS. ESRI Press.

BROWN, P.E., KARESEN, K., ROBERTS, G. E TONELLATO, S. (2000) Blur-Generated Non-Separable Space-Time Models. Journal of the Royal Statistical Society B, 62, no. 4, 847-860.

BROWN, P.J., LE, N. D. E ZIDEK, J. V. (1994) Multivariate Spatial Interpolation and Exposure to Air Pollutants. The Canadian Journal of Statistics, 22, no. 4, 489-509.

CHRISTENSEN, O.F. E RIBEIRO JR, P.J. (2002) geoRglm: A Package for Generalised Linear Spatial Models. R-NEWS, 2, no. 2, 26-28.

CRESSIE, N. E HUANG, H-C (1999) Classes of Nonseparable, Spatio-Temporal Stationary covariance Functions. Journal of the American Statistical Association, 94, no. 448, 1330-1340.

CRESSIE, N.A.C. (1993) Statistical for Spatial Data. Revised Edition. John Wiley & Sons, Inc.

DAMIAN, D., SAMPSON, P.D. E GUTTORP, P. (2001) Bayesian Estimation of Semi-Parametric Non-Stationary Spatial Covariance Structures. Environmetrics, 12, 161-178.

DIGGLE, P. J. (2002) Statistical Analysis of Spatial Point Patterns. London: Edward Arnold. 2nd Edition.

DIGGLE, P.J., MOYEED, R. A. E TAWN, J.A. (1998) Model-Based Geostatistics (with discussion). Applied Statistics, 47, 299-350.

DIGGLE, P.J. E RIBEIRO JR, P.J. (2000) Model Based Geostatistics. Associação Brasileira de Estatística - ABE - 14° SINAPE.

FUENTES, M. E SMITH, R.L. (2000) Modeling Nonstationary Spatial Processes as a Convolution of Local Stationary Processes. Technical Report. North Carolina State University, USA.

GAMERMAN, D. (1997) Markov Chain Monte Carlo - Stochastica Simulation for Bayesian Inference. Chapman & Hall.

GELFAND, A.E., E SMITH, A.F.M. (1990) Sampling-Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association, 85, 398-409.

GREEN, P.J. E SILVERMAN, B.W. (1994) Nonparamentric Regression and Generalized Linear Models: A Roughness Penalty Approach. Chapman and Hall.

HANDCOCK, M.S. E STEIN, M. (1993) A Bayesian Analysis of Kriging. Technometrics, 35, no. 4, 403-410.

HIGDON, D., SWALL, J. E KERN, J. (1999) Non-Stationary Spatial Modeling. In Bayesian Statistics 6 - Proceedings of the Sixth Valencia Meeting, pp.761-768. J.M. Bernardo, J.O. Berger, A.P. Dawid, e A.F.M. Smith, (editors). Clarendon Press-Oxford.

JOURNEL, A.G. E HUIJBREGTS, C.J. (1978) Mining Geostatistics. Academic Press, London.

KIM, H-M., MALLICK, B.K. E HOLMES, C.C. (2002) Analyzing Non-Stationary Spatial Data using Piecewise Gaussian Processes. Tecnhical. Texas.

LE, N.D. E ZIDEK, J.V. (1992) Interpolation with Uncertain Spatial Covariances: A Bayesian Alternative to Kriging. Journal of the Multivariate Analysis, 43, 351-374.

MARDIA, K.V. E GOODALL, C.R. (1993) Spatial-Temporal Analysis of Multivariate Environmental Monitoring Data. In Multivariate Environmental Statistics, pp. 347-386. G.P. Patil e C.R. Rao (editors). Elsevier Science Publishers B.V.

MIGON, H.S. E GAMERMAN, D. (1999) Statistical Inference - an integrated approach. Arnold.

MOLLIÉ, A. (1996) Bayesian Mapping of Disease, em *Markov Chain Monte Carlo in Practice* (eds. W.R.Gilks, S.Richardson, and D. J. Spiegelhalter), pp. 359-379, Chapman & Hall.

NOBRE, A.A. (2003) A Relação entre a Malária e a Chuva no Estado do Pará: Uma Análise Espaço-Temporal. Dissertação de Mestrado, Departamento de Métodos Estatísticos, IM-UFRJ.

O'HAGAN, A. (1994) Kendall's Advanced Theory of Staistics, Volume 2B, Bayesian Inference. John Wiley & Sons, Inc.

OMRE, H., HALVORSEN, K.B. E BERTEIG, V. (1989) A Bayesian Approach to Kriging. In Geostatistics, pp. 109-126. M. Armstrong (editor). Kluwer Academic Publishers.

RIBEIRO, JR., E DIGGLE, P. (2002) Bayesian Inference in Gaussian Model-Based Geostatistics. Geographical and Environmental Modelling, 6, no. 2, 129-146.

RIBEIRO, JR., E DIGGLE, P. (2001) geoR: A Package for Geostatistical Analysis. R-NEWS, 1, no. 2, 15-18.

RIPLEY, B.D. (1981) Spatial Statistics. John Wiley & Sons, Inc.

SAMPSON, P.D. E GUTTORP, P. (1992) Nonparamentric Estimation of Nonstationary Spatial Covariance Structure. Journal of the American Statistical Association, 87, no. 417, 108-119.

SCHMIDT, A.M. (2001) Bayesian Spatial Interpolation of Environmental Monitoring Stations. PhD Thesis – Department of Probability and Statistics, University of Sheffield.

SCHMIDT, A.M. E GELFAND, A.E. (2003) A Bayesian Coregionalization Approach for Multivariate Pollutant Data. Journal of Geophysical Research-Atmospheres , 108.

SCHMIDT, A.M. E O'HAGAN, A. (2003) Bayesian Inference for Nonstationary Spatial Covariance Structures via Spatial Deformations. Journal of the Royal Statistical Society, Series B, 65, 3, 743-758.

SMITH, R.L. (1996) Estimating Nonstationary Spatial Correlations. Technical Report. Cambridge University, UK.

SPIEGELHALTER, D.G., THOMAS, A. E BEST, N.G. (2002) WinBugs Version 1.4 User Manual. Technical Report. Cambridge: Medical Research Council Biostatistics.

STEIN, M.L. (1999) Interpolation of Spatial Data. Springer-Verlag, New York, Inc.

STIDD, C.K. (1974) Estimating the Precipitation Climate. Water Resources Research, 9, 1235-1241.

SUN, W., LE, N.D., ZIDEK, J.V. E BRUNETT, R. (1998) Assessment of a Bayesian Multivariate Interpolation Approach for Health Impact Studies. Environmetrics, 9, 565-586.

Agradecimentos

Alexandra M. Schmidt é Professora Adjunta, Aline A. Nobre é doutoranda (bolsista do CNPq) e Gustavo S. Ferreira é mestrando (bolsista da CAPES) do programa de Pós-Graduação do Departamento de Métodos Estatísticos da Universidade Federal do Rio de Janeiro. Os autores agradecem ao editor Renato Assunção e a um parecerista anônimo, cujas sugestões e comentários melhoraram a apresentação do material aqui exposto.

Abstract

The recent development of stochastic simulation and of computational tools have allowed the use of highly structured models in the many different areas of Statistics. In particular, the area of Spatial Statistics has been receiving a lot of attention from many important researchers and more realistic models have been proposed.

This paper aims to summarize some techniques used to model spatially referenced data. We draw our attention to models for observations taken at different points over a region of interest and, also, to spatial auto-regressive models for areal data. We then use the Bayes' Paradigm to make inference about the parameters in these models. Some characteristics of these two kinds of observations are described and some recent topics of research are pointed. Two examples are analysed, the rainfall in the city of Rio de Janeiro in January of 2002 and the number of cases of malaria in the state of Pará in January of 1997.

Key Words: Anisotropy; areal data; auto-regressive models; Bayes' paradigm; Gaussian process; geographic information system (GIS); geostatistics.

POLÍTICA EDITORIAL

A Revista Brasileira de Estatística - RBEs objetiva promover a Estatística relevante para aplicação em questões sociais, interpretadas, amplamente, para incluir questões educacionais, de saúde, demográficas, econômicas, legais, de políticas públicas e de estatísticas oficiais, entre outras. A revista apresenta artigos num formato que permite fácil assimilação pelos membros da comunidade científica em geral. Os artigos devem incluir aplicações práticas como assunto central. Essas aplicações devem ter conteúdo estatístico substancial. As análises devem ser exaustivas e bem apresentadas, mas o emprego de métodos estatísticos inovadores não é essencial para publicação.

Artigos contendo exposição de métodos são aceitáveis, desde que estes sejam relevantes para as áreas cobertas pela revista, auxiliem na compreensão do problema e contenham interpretação clara das expressões matemáticas apresentadas. A apresentação de aplicações ilustrativas envolvendo dados adequados é requerida. Tratamentos algébricos extensos devem ser evitados.

A RBEs tem periodicidade semestral e publicará, também, artigos escritos a convite e resenhas de livros, bem como artigos abordando os diversos aspectos de metodologias relevantes para órgãos produtores de estatísticas, incluindo:

planejamento de pesquisas;

avaliação e mensuração de erros em pesquisas;

uso e combinação de fontes alternativas de informação; integração de dados;

novos desenvolvimentos em metodologia de pesquisa;

crítica e imputação de dados;

amostragem e estimação;

disseminação e confiabilidade de dados;

análise de dados;

análise de séries temporais;

modelos e métodos demográficos; e

modelos e métodos econométricos.

Todos os artigos submetidos serão avaliados quanto à qualidade e à relevância por dois especialistas indicados pelo Comitê Editorial da Revista Brasileira de Estatística. Os artigos submetidos deverão ser inéditos e não deverão ter sido simultaneamente submetidos a qualquer outro periódico nacional. O processo de avaliação é do tipo duplo cego, isto é, os artigos são avaliados sem identificação da autoria, e os comentários dos avaliadores também são repassados aos autores sem identificação.

INSTRUÇÕES PARA SUBMISSÃO DE ARTIGOS À RBEs

Os artigos submetidos para publicação deverão ser remetidos em três vias (que não serão devolvidas) para:

Renato Assunção
Editor Responsável
Revista Brasileira de Estatística - RBEs
Av. República do Chile 500, 10º andar
Rio de Janeiro – RJ – 20031-170
Tel.: +55 - 21 - 2142 0472
Fax: +55 - 21 - 2142 0039
E-mail: assuncao@est.ufmg.br

Para cada artigo publicado, serão fornecidas gratuitamente 20 separatas.

Instruções para preparo de originais

Os originais entregues para publicação devem obedecer às seguintes normas:

1. A primeira página do original (folha de rosto) deve conter o título do artigo, seguido do(s) nome(s) completo(s) do(s) autor(es), indicando-se, para cada um, a filiação e endereço para correspondência. Agradecimentos a colaboradores e instituições, e auxílios recebidos devem figurar, também, nesta página;

2. A segunda página do original deve conter resumos em português e em inglês (*Abstract*), destacando os pontos relevantes do artigo. Cada resumo deve ser datilografado seguindo o mesmo padrão do restante do texto, em um único parágrafo, sem fórmulas, com no máximo 150 palavras;

3. O artigo deve ser dividido em seções numeradas progressivamente, com títulos concisos e apropriados. Todas as seções e subseções devem ser numeradas e receber título apropriado;

4. A citação de referências no texto e a listagem final das referências devem ser feitas de acordo com as normas da ABNT;

5. As tabelas e gráficos devem ser precedidos de títulos que permitam perfeita identificação do conteúdo. Devem ser numeradas seqüencialmente (Tabela 1, Figura 3, etc.) e referidas nos locais de inserção pelos respectivos números. Quando houver tabelas e demonstrações extensas ou outros elementos de suporte, podem ser empregados apêndices. Os apêndices devem ter título e numeração, tais como as demais seções do trabalho;

6. Gráficos e diagramas para publicação devem ser incluídos nos arquivos com os originais do artigo, sempre que possível. Quando isto não ocorrer, devem ser traçados em papel branco, com nitidez e boa qualidade, para permitir que a redução seja feita mantendo qualidade. Fotocópias não serão aceitas. É fundamental que não existam erros, quer no desenho, quer nas legendas ou títulos; e

7. Serão preferidos originais processados pelo editor de texto *Word* for *Windows*.