UNIVERSIDADE FEDERAL DE LAVRAS

JOÃO MARCOS LOUZADA

TESTE DE NORMALIDADE MULTIVARIADA EM GEOESTATÍSTICA UTILIZANDO BOOTSTRAP

LAVRAS – MG

2011

JOÃO MARCOS LOUZADA

TESTE DE NORMALIDADE MULTIVARIADA EM GEOESTATÍSTICA UTILIZANDO BOOTSTRAP

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Orientador

Dr. Marcelo Silva de Oliveira

LAVRAS - MG

2011

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca da UFLA

Louzada, João Marcos.

Teste de normalidade multivariada em Geoestatística utilizando bootstrap/ João Marcos Louzada. – Lavras: UFLA, 2011.

167 p. : il.

Tese (Doutorado) – Universidade Federal de Lavras, 2011.

Orientador: Marcelo Silva de Oliveira.

Bibliografia.

1. Estimação de modelos de predição espacial. 2. Modelagem de processos estocáticos. 3. Semivariograma. 4. Teste Gaussiano para geoestatística. I. Universidade Federal de Lavras. II. Título.

CDD–519.54

JOÃO MARCOS LOUZADA

TESTE DE NORMALIDADE MULTIVARIADA EM GEOESTATÍSTICA UTILIZANDO BOOTSTRAP

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 8 abril de 2011.

Prof. Dr. Ednaldo Carvalho Guimarães	UFU
Prof. Dr. Marcelo de Carvalho Alves	UFMT
Prof. Dr. Renato Ribeiro de Lima	UFLA
Prof. Dr. João Domingos Scalon	UFLA

Prof. Dr. Marcelo Silva de Oliveira
Orientador

LAVRAS - MG

2011

A maravilhosa sabedoria dos desígnios divinos

Ó profundidade da riqueza, tanto da sabedoria como do conhecimento de Deus! Quão insondáveis são os seus juízos, e quão inescrutáveis, os seus caminhos! Quem, pois conheceu a mente do Senhor? Ou quem foi o seu conselheiro? Ou quem primeiro deu a Ele para que lhe venha a ser restituído? Porque Dele, e por meio Dele e para Ele são todas as coisas. A Ele, pois a glória eternamente. Amém.

Romanos 11: 33-36

A minha esposa, Andreia, pelo apoio e carinho; meus queridos filhos, Tarcylla e Thales, meus irmãos e aos meus pais

DEDICO

AGRADECIMENTOS

Ao Deus de Abraão, de Isaque e de Jacó eu louvo por mais essa vitória em minha vida.

Aos meus queridos pais, Veredino e Jandyra, pelo exemplo de fé, incentivo e amor perfeito.

À Andreia, minha querida esposa e companheira de todas as horas, pela valentia, dedicação, compreensão, ternura e insondável amor. "Nem mesmo as muitas águas não poderiam ..."

Aos meus queridos filhos, Tarcylla e Thales, presentes da Eternidade, por serem minha fonte de inspiração e motivação para superar os momentos de desafios e proseguir nessa jornada sempre de cabeça erguida.

Ao meu orientador, Dr. Marcelo Silva de Oliveira, pelo aprendizado, confiança e deveras sapiência na condução deste trabalho.

À Coordenadoria para o Aperfeiçoamento de Pessoal do Nível Superior (CAPES), pela concessão de uma bolsa de estudo, por meio do programa PCID-TEC.

À Universidade Federal de Lavras e ao Departamento de Ciências Exatas, bem como seu capacitado corpo docente, pela oportunidade de aprendizado e pelo incentivo à busca do ser e obter.

Ao grande amigo e irmão Veredino e minha cunhada Adriana, pelo incentivo, apoio e aconselhamentos construtivos.

Aos colegas do programa de pós-Graduação em Estatística e Experimentação Agropecuária da UFLA, pela convivência e troca de ideias.

Em especial ao amigo Gerson, grande parceiro nos estudos, pelas muitas experiências vivenciadas e conhecimentos compartilhados que fizeram a diferença.

Ao colega de doutorado Enio, pelo apoio moral e suas brilhantes ideias que contribuíram significativamente para o êxito do trabalho.

Ao IFES-Campus de Itapina, por ter me concedido a oportunidade de realizar o doutorado.

Ao irmãos da Igreja Cristã Maranata de Lavras, MG, pelo apoio, carinho, convívio e comunhão.

Ao casal Lindomar e Gláucia, pelo incentivo, companherismo e grande apoio.

Ao meu cunhado, Ivan, pelo inestimável apoio nesses quatro anos de doutorado.

Aos membros da banca examinadora: Dr. Ednaldo Carvalho Guimarães, Dr. Marcelo de Carvalho Alves, Dr. Renato Ribeiro de Lima e Dr. João Domingos

Scalon, pelas valiosas contribuições e aprimoramento deste trabalho.

Aos funcionários do DEX e da biblioteca, pela eficiência e boa vontade.

"I never thought that others would take them so much more seriously then I did".

Albert Einstein

RESUMO

Este trabalho foi realizado com o principal objetivo de elaborar um teste de normalidade multivariada para dados espacialmente contínuos, cognominado de Teste Gaussiano para Geoestatística (TGGeo), com o auxílio da metodologia bootstrap. O bootstrap foi fundamental para se construir os limites de confiança percentil, a partir do semivariograma de nuvens. O TGGeo é baseado no ajuste do semivariograma experimental e, basicamente, é utilizado para verificar a normalidade multivariada em dados de Geoestatística pela coincidência satisfatória das modelagens do semivariograma baseadas nos critérios de quadrados mínimos e de máxima verossimilhança. A normalidade é verificada fazendo-se uso da equivalência dos métodos da MV e dos QM na inferência, sob normalidade pelo fato de o critério da máxima verossimilhança ser idêntico ao método dos quadrados mínimos para modelar a dependência espacial. O teste é formulado e a decisão é tomada por meio de inspeção gráfica e avaliando o valor de plausibilidade (valor-pl), dado por δ. Esse valor crítico (δ) indica que percentagem das unidades preditas ($\gamma(h,\theta)$), estimadas por máxima verossimilhança, estão dentro dos limites com $100 \times (1 - \alpha)$ de confiança. Em geral, os resultados mostraram que o TGGeo foi robusto quanto aos desvios de processos gaussianos e, comportou-se de modo razoável em presença de dados com distribuição gaussiana. Para avaliar o teste proposto, foram utilizados conjuntos de seis dados simulados, com diferentes parâmetros de covariância, bem como três conjuntos de dados reais, em que o TGGeo mostrou-se capaz de identificar campos aleatórios gaussianos. Esse mesmo princípio bootstrap possibilitou a construção do semivariograma de quantis para avaliar as incertezas associadas ao procedimento de modelagem do semivariograma experimental e um exemplo de aplicação em dados reais justificou o uso dessa metodologia como uma vantajosa opção para a análise variográfica. Foi também proposto um critério de ponderação, $W_{N/h}$, para modelar o semivariograma por meio dos quadrados mínimos. Os pesos são calculados pela razão entre o número de pares de pontos $N(h)$ e suas respectivas potências das distâncias, h.

Palavras chave: Geoestatística. Teste de normalidade. Semivariograma. Teste Gaussiano para Geoestatística. Bootstrap.

ABSTRACT

This work was undertaken with the main objective of designing a multivariate normality test of spatially continuous data, nicknamed Gaussian Test for Geostatistics (TGGeo) with the aid of bootstrap methodology. Bootstrap was fundamental to construct the percentile confidence limits from the semivariogram cloud. TGGeo is based upon the adjustment of the experimental semivariogram and basically, it is utilized to verify the multivariate normality in data of Geostatistics through the satisfactory coincidence of the modeling of the semivariogram based on the criteria of least squares and maximum likelihood. Normality is verified by making use of the equivalency of the methods ML and of the least squares in the inference under normality for the fact of the criterion of maximum likelihood being identical to the least square method to modeling the spatial dependence. The test is formulated and the decision is made by means of graphical inspection and by evaluating the plausibility value (valor-pl), given by δ . That critical value (δ) indicates that the percentage of the predicted units ($\gamma(h;\theta)$), estimated by maximum likelihood, are within the limits with $100 \times (1-\alpha)$ of confidence. In general, the results showed that TGGeo was robust as to the deviations of Gaussian processes and, behaved itself in a reasonable manner in the presence of data with Gaussian distribution. To evaluate the propose test, sets of six simulate data, with different covariance parameters as well as three sets of real data in which TGGeo proved capable of identifying the random Gaussian fields. That same bootstrap principle made the construction of the quantile semivariogam to evaluate the uncertainties associated with the procedure of modeling of the experimental semivariogram possible and an example of application in real data accounted for the use of that methodology as an option advantageous to the variographic analysis. A weighing criterion, $W_{N/h}$, was also proposed to model the semivariogram by means of least squares. The weights are calculated by the ration between the number of pairs of points $N(h)$ and their respective powers of the distances, h.

Key words: Geostatistics. Normality test. Semivariogram. Gaussian test for Geostatistics. Bootstrap.

SUMÁRIO

1 INTRODUÇÃO 12
2 REFERENCIAL TEÓRICO 17
2.1 Processos estocásticos: uma abordagem voltada para dados espaciais 17
2.1.1 Conceito de processos estocásticos: construção do modelo probabilístico para Geoestatística 20
2.1.2 Formalização do modelo matemático 29
2.2 Geoestatística 31
2.2.1 Contextualização e conceitos 32
2.3 Estacionaridade do processo estocástico e tendência nos dados . 40
2.3.1 Estacionaridade dos dois primeiros momentos 43
2.3.2 Semivariograma 47
2.3.2.1 Semivariogramas de nuvens: dissimilaridade versus distância . . 48
2.3.2.2 Semivariograma experimental: semivariância média 49
2.4 Método de ajuste do semivariograma 54
2.4.1 Estimação por máxima verossimilhança 55
2.4.2 Estimação por máxima verossimilhança restrita 57
2.4.3 Método dos quadrados mínimos e o semivariograma 60
2.5 Modelos teóricos de semivariogramas 63
2.5.1 Semivariograma com patamar ou modelos de transição 64
2.5.2 Modelos de semivariogramas sem patamar 71
2.6 Outros modelos gerais de semivariogramas 74
2.7 Propriedades intrínsecas do padrão espacial: isotropia, anisotropia 77
2.7.1 Isotropia e anisotropia 78
2.8 Metodos bootstrap 80
2.8.1 Visão geral 80
2.8.2 Fundamentos básicos 81
2.8.3 Teoria bootstrap: uma ideia genial 83
2.8.4 O princípio *plug-in* 85
2.8.5 Estimativa do erro padrão bootstrap 89
2.9 Limites de confiança percentil bootstrap 90
2.10 Sumário 91
2.11 Notas históricas 93
3 MATERIAIS E MÉTODOS 96

3.1 Informações a priori 96
3.2 Bootstrapping: um exemplo de simulação 98
3.2.1 Aspectos computacionais gerais e simulação estocástica 104
4 RESULTADOS E DISCUSSÕES 111
4.1 Resultados metodológicos 111
4.1.1 Trabalhando os dados: definições e nomenclaturas 111
4.1.2 Usando o bootstrap na modelagem do semivariograma 114
4.1.3 Teste gaussiano para Geoestatística (TGGeo) 126
4.1.4 Cálculo do valor de plausibilidade 132
4.1.5 Critério de ponderação para ajustar o semivariograma por quadrados mínimos 134
4.1.6 Semivariograma de quantis bootstrap 137
4.2 Teste de normalidade para Geoestatística 142
4.2.1 Campos aleatórios gaussianos sob correlação esférica 142
4.2.2 Campos aleatórios gaussianos sob correlação exponencial 143
4.2.3 Campos aleatórios gaussianos sob correlação gaussiana 144
4.2.4 Campos aleatórios não gaussianos simulados sob correlação esférica 148
4.2.5 Campos aleatórios não gaussianos simulados sob correlação exponencial 149
4.2.6 Campos aleatórios não gaussianos simulados sob correlação gaussiana 150
4.2.7 Teste gaussiano para Geoestatística: uma aplicação em dados reais 154
5 CONCLUSÕES 160
REFERÊNCIAS 162

1 INTRODUÇÃO

Esta pesquisa busca contribuir efetivamente para aperfeiçoar os métodos da Geoestatística no que diz respeito à modelagem de processos estocásticos regionalizados. Incrementaram-se ou incorporaram-se nos procedimentos analíticos da modelagem do semivariograma os benefícios das técnicas de reamostragem via método bootstrap. Nesse contexto, espera-se que essa abordagem possa oferecer vantagens àqueles que se utilizam da modelagem variográfica como meio de predição espacial de processos estocásticos. Tais inovações possibilitaram a construção de valiosos intervalos de confiança para os parâmetros de covariância espacial (efeito pepita, patamar e alcance) e viabilizou, também, a elaboração de semivariogramas de quantis, dando ao pesquisador mais poder de análise, tornando possível avaliar as incertezas associadas na estimação do semivariograma experimental, como também no seu modelo teórico ajustado. O processo de estimação do semivariogram envolve dois níveis de incertezas: o primeiro na estimação do semivariograma experimental e o segundo refere-se à utilização de um método convencional qualquer para suavizar as estimativas do semivariograma experimental por meio de um modelo ajustado.

Em geral, com a metodologia aqui apresentada, podem-se obter resultados mais precisos na fase de caracterização do padrão de dependência espacial (modelagem do semivariograma). Nesse sentido, é de suma relevância destacar aqueles casos em que os ensaios de campo (levantamento de dados espaciais) são realizados em processos de grandes interesses econômicos, porém, essencialmente complexos e em sua maioria de alto custo, como é o caso das amostragens em reservas de petróleo, inventários florestais, temas geológicos, agricultura de precisão e nas áreas de mineração. Destacam-se esses campos da ciência, particularmente,

mas, na verdade, trata-se de um problema cotidiano vivido pelos geoestatísticos: o fato de lidar com as incertezas inerentes à natureza do processo estocástico e que também se fazem presentes na estimação do semivariograma.

Desse modo, o pesquisador terá à disposição várias ferramentas de análise geoestatística baseada em computação intensiva, as quais serão recursos valiosos na tomada de decisão quanto à estimação de modelos de predição espacial e que ainda são pouco difundidas entre os geoestatísticos — mesmo sendo a métodologia bootstrap não recente e já bastante aplicada em inferências da Estatística clássica. Mas, no tocante às práticas de Geoestatística, ainda é uma técnica pouco explorada, e a abordagem bootstrap que se apresenta nesta tese é apenas a "ponta visível do *iceberg*" porque, ainda, grandes são os desafios para adequá-la ou ajustá-la mais simplesmente ao universo geoestatístico.

Há casos em que os fenômenos possuem dependência espacial, mas suas estruturas são difícieis de ser perfeitamente capturadas pelo estimador padrão de semivariâncias e, portanto, não são identificadas claramente pela inspeção variográfica. Em outras palavras, o alcance, o patamar e o tipo de modelo não estão bem definidos (ou nítidos), o que gera dúvidas quanto à qualidade dos modelos ajustados. Nessas situações, o bootstrap oferece ao pesquisador a possibilidade de gerar intervalos de confiança para todo o procedimento de estimação do semivariograma, o que dá ao geoestatístico mais capacidade e confiabilidade em suas análises - é importante ressaltar que o semivariograma é a "menina dos olhos"da Geoestatística e dele depende o "coração"da mesma, a krigagem, para realizar predições precisas e confiáveis.

O bootstrap é uma ferramenta interessante para esse estudo, porém, ainda avança muito timidamente em Geoestatística. Nesta área tem-se um número reduzido de trabalhos com bootstrap. Portanto, esta tese é princípio daquilo que

se acredita ser um vasto e promissor laboratório de pesquisa. Nessa instância, utilizou-se, modestamente, a capacidade desse método estimar parâmetros sem requerer nenhuma suposição além da independência dos dados, sendo nesse sentido não paramétrico.

Muitos métodos estatísticos tradicionais repousam sob o uso da distribuição normal para os dados. Porém, como a normalidade exata não é tão comum, tem-se aí uma divergência com a demanda teórica. Sabe-se que os procedimentos inferenciais usando a distribuição t são robustos quanto aos desvios da normalidade e, portanto, muito úteis na prática. Contudo, exceto para grandes amostras, não se podem usar intervalos de confiança t e fazer testes se os dados são fortemente assimétricos. Por causa desses problemas abordados, o bootstrap — um método de reamostragem baseado nos dados — está se tornando cada vez mais popular como uma metodologia estatística. Na verdade, com o procedimento bootstrap, a distribuição empírica do processo pode ser construída, sem se importar com a necessidade de dados normais ou grandes amostras e também traz a libertação de fórmulas.

Voltando à Geoestatística, ela disponibiliza para a ciência um poderoso conjunto de métodos para modelagem de dados espaciais. Uma abordagem ainda incipiente e pouco explorada é a Geoestatística baseada em modelos, a qual requer a especificação da distribuição conjunta dos dados espaciais, que são modelados como uma normal multivariada (DIGGLE; RIBEIRO JUNIOR, 2007). O fato de essa metodologia ter baixa difusão no Brasil, além de não ser de domínio da grande maioria dos geoestatísticos, pode estar relacionado às seguintes questões: i) a Geoestatística é bastante utilizada por áreas bem aplicadas e a abordagem baseada em modelos exige um conhecimento mais apurado da Estatística clássica e ii) em geral, suas modernas ferramentas analíticas (inferência baseada em verossimilhança,

métodos bayesianos, modelos mistos generalizados, etc.) ainda não foram implementadas nos programas de computadores mais comumente usados para modelagem geoestatística e, geralmente, estão disponíveis no programa estatístico R que não é contemplado pelo usuário não-estatístico, devido à sua interface ser menos amigável que as dos programas comerciais.

No contexto da Geoestatística, sabe-se que esses métodos têm evoluído bastante nessa última década. Houve também considerável avanço computacional dos mesmos, em que boa parte desses métodos já tem sido implementada em diversos programas de computador. Entretando, acredita-se que esse volume de informações seja mais rapidamente implementado em programas de código aberto que nas empresas particulares, obviamente. Como exemplo, destaca-se o projeto R (R DEVELOPMENT CORE TEAM, 2009), uma iniciativa de código livre que é fomentada por desenvolvedores de toda a comunidade científica mundial, ao contrário das empresas privadas que contam com um número restrito desses profissionais. Contudo, devido ao fato de as pesquisas científicas em Geoestatística beseada em modelos serem relativamente incipientes, embora tenham-se obtidos preciosos avanços teóricos e computacionais, como já mencionado, torna-se evidente a grande demanda de pequisas profícuas e de vanguarda nesse campo da ciência (DIGGLE; RIBEIRO JUNIOR, 2007; PEBESMA, 2004; SCHLATHER, 2010; RUE; TJELMELAND, 2002).

Considerando a classe de problema (i) acima e o fato de a Geoestatística baseada em modelo repousar sobre a suposição de que o processo estocástico subjacente obedece a uma distribuição de probabilidade, sendo a mais natural e conveniente, a distribuição normal multivariada para variáveis regionalizadas, um questionamento é de fundamental importância: o problema de testar a normalidade multivariada em Geoestatística. Este é o problema de pesquisa desta tese.

Em decorrência, portanto, desse problema apresentado, esta pesquisa foi realizada com os seguintes objetivos:

- adaptar a metodologia bootstrap à realidade variográfica e elaborar um teste para normalidade multivariada em Geoestatística, isto é, construir um teste de normalidade multivariada para dados espacialmente contínuos com o auxílio do bootstrap;
- construir semivariogramas e intervalos de confiança baseados em quantis bootstrap.

2 REFERENCIAL TEÓRICO

2.1 Processos estocásticos: uma abordagem voltada para dados espaciais

A maneira mais simples de se compreender o conceito de processos estocásticos é pensá-lo como uma coleção de variáveis aleatórias $Z(x_i)$, indexadas segundo um conjunto de índices $i = 1, 2, \ldots, n$. Outros autores preferem usar a terminologia famílias de variáveis aleatórias e as denominam de funções aleatórias, com o mesmo objetivo (WEBSTER; OLIVER, 2001). Essa família, compreendida como uma coleção de objetos, é gerada por um mecanismo aleatório subjacente a um dado fenômeno que, provavelmente, está ocorrendo dentro de um espaço a ser estudado. Em outras palavras, há a ocorrência de um ou mais fenômenos que podem ser tratados e postulados probabilisticamente, sendo quantificados por meio de realizações de variáveis aleatórias concernentes a processos que se desenvolvem no tempo ou no espaço ou em ambos. De fato, existem muitas aplicações práticas de probabilidades interessadas em processos aleatórios, quer seja no tempo e/ou espaço, os quais podem ser discretos ou contínuos (CRESSIE, 1993).

Segundo Wackernagel (2003), os dados fornecem informações sobre as variáveis regionalizadas, as quais são simplesmente funções $z(x)$ cujo comportamento deseja-se caracterizar em uma dada região $\mathcal{R}$ de um espaço contínuo. Journel e Huijbregts (1991) afirma que a definição variável regionalizada como uma variável distribuída no espaço é puramente descritiva e não envolve nenhuma interpretação probabilística e que do ponto de vista matemático, a variável regionalizada é simplesmente uma função $f(x)$ a qual toma um valor em todo ponto x de coordenadas (x_i, y_j, z_k) no espaço tridimencional. Em um modelo probabilístico, a variável regionalizada $z(x)$ é considerada uma realização de uma função aleatória $Z(x)$. Assim, um conjunto de dados é uma amostra de uma particu-

lar realização $z(x)$ de $Z(x)$. Conforme Wackerganel (2003), a vantagem dessa abordagem é que se deve somente tentar caracterizar as simples características da função aleatória $Z(x)$ e não aquelas particulares realizações $z(x)$.

Dessa forma, cada valor medido no conjunto de dados é visto como um valor regionalizado, mensurado no domínio de $\mathcal{R}$ e pode ser considerado como um resultado de algum mecanismo aleatório. De acordo com Wackernagel (2003), esse mecanismo é formalmente denominado de variável aleatória e um valor amostrado $z(x_i)$ representa um sorteio da variável $Z(x_i)$.

Wackernagel (2003) traz uma interessante explanação sobre a ideia do modelo de função aleatória que estabelece ou determina os dados sob duas diferentes perspectivas. Um aspecto é o fato de os valores dos dados serem originados de um ambiente físico (espacial ou temporal) e são de alguma maneira dependente de suas localizações na região: eles são regionalizados. O segundo aspecto é que os valores da amostra regionalizada $z(x_i)$ não podem, geralmente, ser modelados como um fenômeno determinístico, simplesmente, isto é, o mecanismo é considerado como aleatório. Com esse modelo, os valores dos dados são vistos como realizações (ou resultados) do mecanismo aleatório e esses dois aspectos juntos, de regionalização e aleatorização produzem o conceito de uma de função aleatória. Um boa ilustração desse conceito encontra-se na Figura 2.1.

Grandezas estocásticas unidimensionais são medições no tempo, como ocorre em processos aleatórios típicos de aplicações em séries temporais. Já os processos aleatórios espaciais disponíveis em uma região de interesse, $\mathcal{R}$, podem ser representados pelo espaço euclidiano bidimensional, em que o fenômeno é comumente amostrado e posteriormente submetido à análise estatística (geralmente de cunho preditivo), a qual deve ser validada por meio de um suposto modelo de probabilidade. Realmente, o emprego do nome processo estocástico é adequado

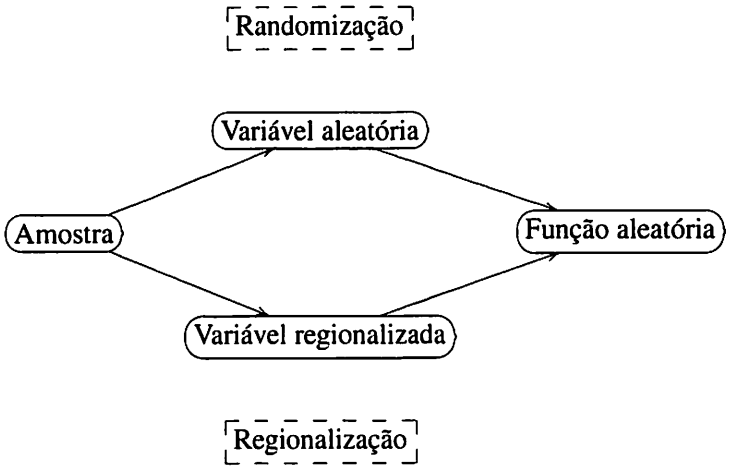


Figura 2.1 O modelo de função aleatória
Fonte Wackernagel (2003)

por se tratar de uma coleção ou familia de variáveis aletórias. Atualmente, no contexto da Geoestatística baseada em modelos, também vem sendo empregada a terminologia campos aleatórios (DIGGLE; RIBEIRO JUNIOR, 2007). Assim, os nomes processos estocásticos, processos aleatórios, funções aleatórias e campos aleatórios possuem o mesmo significado. A palavra estocástico é proveniente de um radical Grego "$\sigma\tau\grave{o}\chi os$", ou da sua forma adjetivada "$\sigma\tau o\chi\alpha\sigma\tau\iota\kappa o\xi$", que significa almejar ou adivinhar (WHITT, 1986).

Whitt (1986) declara que, historicamente, o desenvolvimento teórico de processos estocásticos não está bem delineado, trazendo relatos imprecisos e, consequentemente, conduzindo a informações não-confiáveis sobre o tema. Sua origem está pulverizada por toda literatura, porém, geralmente limitada a trabalhos específicos e de áreas mais aplicadas, tais como a Física e Economia, por exemplo. Além disso, a maioria das abordagens está voltada para aplicações em séries temporais econométricas. Segundo Whitt (1986), embora o desenvolvimento teórico de processos estocásticos tenha iniciado por meio dos físicos-estatísticos

(Gibbs, Boltzmann, Poincaré, Smoluchowski, Langevin) e, depois, pelo trabalho pioneiro sobre movimento browniano (Einstein, Wiener, Lévy), a fundamentação teórica foi construída por Doob, Komolgorov e outros. Nesse contexto, destaca-se ainda o pioneiro trabalho de Doob sobre *Markov processes*[1] and martingales[2] que é realmente um marco em se tratando de processos estocásticos. Assim, esses desenvolvimentos culminaram na grande gama de tópicos variando desde as teorias gerais mais recentes de processos estocásticos a um variado número de aplicações em diversos campos (KANNAN, 1979).

2.1.1 Conceito de processos estocásticos: construção do modelo probabilístico para Geoestatística

É lícito supor a existência de vários fenômenos ocorrendo dentro de uma área $\mathcal{R}$ cujos comportamentos estão diretamente associados ou dependentes de suas particulares localizações espaciais, isto é, esses elementos comumente possuem qualidades bastante regionalizadas, no sentido de que o processo não ocorre de forma desconexa no contexto espacial e, portanto, não se imagina obter quantidades próximas muito dissimilares — espera-se que os dados tomados próximos uns dos outros apresentem certa familiaridade (ISAAKS; SRIVASTAVA, 1989). Desse modo, o fenômeno é regido por um subjacente processo que define sua forma natural, configuração espacial e padrão estrutural característico, os quais

[1]Um processo aleatório, $Z(x)$, é processo de Markov se a probabilidade condicional de qualquer evento futuro, dado qualquer evento passado e o presente, é independente do evento passado e depende apenas do estado presente - também denominado processo sem memória, uma vez que o passado é "esquecido".

[2]Conforme Kannan (1979), o termo martingale é na verdade uma sigla francesa empregada em sitema de aposta (risco), em que o jogador aumenta (ou dobra) as apostas até vencer o jogo. Mas, a teoria martingale é uma poderosa ferramenta em teoria de probabilidade e não se restringe apenas a estratégias de jogos, a saber: um processo estocástico $\{X_n, n \geq 0\}$ é dito ser um matingale em relação a um processo $\{Y_n, n \geq 0\}$ se, para todo $n \geq 0$, $E[\mid X_n \mid] < \infty$ e $E[X_{n+1} \mid Y_0, \ldots, Y_n] = X_n$.

estão associados ao conceito de vizinhança. Essa construção racional parte do princípio de que esse tipo de processo está organizado espacialmente na natureza e, portanto, o mesmo repousa sobre uma lei "maior" que define claramente o seu próprio comportamento, bem como esses fenômenos se relacionam uns com os outros. Mesmo sob essas condições razoáveis, em alguns casos, eles podem não se relacionar entre si e, nesse caso, não se deve esperar que qualquer valor de dado avaliado tenha algum tipo de influência sobre os dados circunvizinhos, quer seja essa quantidade disponível pertencente a mesma característica ou de características distintas.

De modo geral, dizer que um fenômeno é espacialmente estruturado implica em considerar que, a princípio, as características estudadas possuem dependência estatística gerada na vizinhança de qualquer unidade espacial (ponto amostrado), o que garante o seguinte princípio quantitativo: os elementos amostrais mais próximos (adjacentes) tendem a ser similares, mas essa similaridade torna-se menor com o gradativo aumento da distância entre esses elementos do mesmo fenômeno. De fato, quanto maior for a distância entre os pares de elementos amostrados, espera-se que menor seja a similaridade entre os valores medidos (WACKERNAGEL, 2003). Note que esse importante aspecto está diretamente relacionado com a variabilidade espacial do processo, cujos padrões espaciais podem ser razoavelmente modelados e mapeados pelos termos da Geoestatística, formando, assim, o que se denomina de padrão de dependência espacial. Por sua vez, a medida de similaridade-dissimilaridade está diretamente relacionada com o conceito de autocorrelação espacial, análogo ao coeficiente de correlação desenvolvida por Pearson, por meio do qual se pode quantificar o padrão espacial (característica peculiar) de cada fenômeno que se comporta como tal (SCHABENBERGER; GOTWAY, 2005; WEBSTER; OLIVER, 2001; ISAAKS; SRIVASTAVA, 1989).

A autocorrelação é um dos métodos mais antigos de se estimar a dependência no espaço de amostras vizinhas. Modelos para análise de autocorrelação foram desenvolvidos principalmente por geógrafos ingleses no início da década de 1970 e aplicados à Biologia de Populações a partir dos trabalhos de Jumars, Thlstle e Jones (1977) e Sokal (1978a, 1978b). Mais adiante o conceito de autocorrelação será tratado em detalhes.

Agora, faz-se necessário abordar como a Estatística Espacial trata todo esse comportamento inerente às variáveis espacialmente estocásticas. E, ainda, como se deve elaborar um suposto modelo que seja capaz de validar o suporte teórico geralmente utilizado pelos métodos da Estatística Espacial, com o objetivo de inferir sobre a população, com base em amostras discretas retiradas a partir de processos contínuos e espacialmente correlacionados. Uma forma bastante simples de se responder essa demanda é considerar que esses dados estão sendo continuamente gerados a partir de um processo regionalizado, desconhecido, mas que certamente está agindo ocultamente sobre a área $\mathcal{R}$. Note que o processo gerador de dados, ocorrendo na área, gera uma população altamente estruturada e que, devido a esse fato, torna-se irrelevante a adoção do princípio da aleatorização nos procedimentos de tomada de amostras autocorrelacionadas. Esse procedimento é de suma importância na composição do modelo da Estatística convencional que repousa sobre a suposição de independência entre os indivíduos. Porém, o mesmo não faz sentido nos casos em que se deseja justamente caracterizar ou modelar esse padrão espacial existente e, portanto, torna-se um contra-senso tentar aleatorizar essa estrutura natural do fenômeno (CRESSIE, 1993; WALLER; GOTWAY, 2004).

Assim, o procedimento de coleta dos dados deve ser determinístico, de forma que os pontos escolhidos possam cobrir satisfatoriamente toda a área de in-

teresse e, além disso, essa forma de amostragem não interfere no comportamento natural das informações disponíveis (OLIVEIRA, 1991). Porém, existem problemas quanto à configuração e ao espaçamento dos pontos amostrais que devem ser observados, porém, não são objetos de pesquisa desta tese e que será abordado apenas para subsidiar a argumentação. Webster e Oliver (2001, p. 85) trazem um boa discussão sobre os problemas de amostragem, geralmente, encontrados em Geoestatística.

Outra questão de suma importância demanda o conhecimento sobre o modelo que dá sustentabilidade e plausibilidade à metodologia empregada para se tratar com fenômenos que possuem uma distribuição espacial continuamente autocorrelacionada (d.e.c.a). Müller (2007) afirma que duas características mais peculiares de delineamentos espaciais são que as observações não podem ser replicadas instantaneamente e que elas são usualmente correlacionados sobre as localizações. Evidentemente, é de se esperar que a amostra levantada dentro da região $\mathcal{R}$ conserve ou reflita as mesmas propriedades inerentes à população que a originou, segundo um planejamento amostral aceitável. A propósito, na Geoestatística, desde o planejamento amostral até a finalização da amostragem no campo, tem-se, em geral, um procedimento bastante interativo, exigindo paciência e muita perícia do especialista para efetuar uma amostragem capaz de detectar satisfatoriamente a dependência espacial. Ainda, a concretização geral da malha amostral regular ou irregular depende fortemente da natureza do fenômeno e, obviamente, do custo operacional da mesma que influencia diretamente no número de pontos a serem amostrados, impactando na qualidade da predição espacial. Por isso, a decisão pela amostra ideal[3] em Geoestatística requer um conhecimento requintado

[3]Entende-se por amostra ideal como sendo a combinação otimizada entre o número de pontos amostrados e seus espaçamentos entre si, de forma que esse conjunto de dados espaciais selecionados seja a "melhor" coleção de variáveis de todas as possíveis realizações dentro da área $\mathcal{R}$.

do pesquisador para conduzir o processo de amostragem (se necessário, com reiterações) até obter uma amostra que configure claramente a estrutura espacial do fenômeno (MÜLLER, 2007). Essa fase é deveras importante, pois uma amostra não-representativa do processo, definitivamente, por ser inábil para absorver a dependência espacial, pode induzir a conclusões equivocadas de ausência de padrão espacial, quando, na verdade, o mesmo existe de fato. Assim, até o momento, deve estar claro que os dados usuais nas análises de Geoestatística carregam, obviamente, a herança da população e, consequentemente, não se pode tratá-los como oriundos de processos randômicos. Tal conceito é originalmente inaceitável e viola o princípio harmonioso da natureza populacional aqui definida, bem como o princípio fundamental da 1ª lei da Geografia, que afirma: "Todas as coisas estão relacionadas umas com as outras, mas aquelas mais próximas são mais semelhantes que mais distantes" (TOBLER, 1969, p. 7).

Uma vez entendido que a metodologia da Geoestatística auxilia o pesquisador a compreender o comportamento da variabilidade espacial dos fenômenos de interesse científico, mediante sua habilidade para tratar dados que possuem uma d.e.c.a., então é se que faz necessário introduzir os conceitos de processos estocásticos multivariados e, posteriormente, os univariados.

Segundo Chilès e Delfiner (1999), a informação disponível sobre um fenômeno natural é raramente limitada aos valores assumidos por uma única variável sobre um conjunto de pontos amostrais. Em aplicações de ciências da Terra existem pelo menos quatro fontes adicionais de informações — a maioria de estudos reais envolve mais de uma variável. São elas:

1. valores numéricos de outras variáveis possivelmente encontrados em outras localizações;

2. valores numéricos da mesma variável em outros pontos de tempo;

3. relacionamentos físicos entre variáveis;

4. conhecimento (ou experiência) de aplicações específicas.

Continuando a estudar esse conceito, imagine uma área de estudo $\mathcal{R}$ em que, supostamente, existem vários fenômenos ocorrendo simultaneamente, e que todos esses fenômenos são de potencial interesse do pesquisador. Em outras palavras, trata-se de um processo multivariado no seu estado lantente, atuando sobre uma área $\mathcal{R}$ (WACKERNAGEL, 2003). Todavia, nesta tese, seria conveniente usar uma nova terminologia, a saber: complexo espacialmente estocástico. Essa nomenclatura é justificada porque a área de pequisa, $\mathcal{R}$, pode ser vista como uma composição de vários elementos (fenômenos estocásticos) ou um conjunto de variáveis aleatórias que possuem algum tipo de ligação entre si ou não e, também, porque cada particular variável aleatória sofre uma autodependência quando se processam medições da própia variável defasada no espaço euclideano, isto é, há uma dependência espacial da variável com ela mesma em função da distância. Note que esse processo se diferencia daquele usual da Estatística multivariada convencional, em que não existe dependência entre as unidades amostrais, mas somente há dependência entre as variáveis. Assim, o termo complexo é usado no sentido de que existe uma depedência espacial tanto entre as característica estudadas quanto dentro da própria característica e, portanto, ambas situações podem influenciar, significativamente, o comportamento de um elemento amostral. Para melhor elucidar a abordagem multivariada em Geoestatística (com exemplos de aplicação analisados no programa R) e ratificar a complexidade desse enfoque, recomenda-se a leitura do artigo de Pebesma (2004). Esse autor declara que a Geoestatística multivariada envolve as predições (ou simulações) simultâneas de variáveis múltiplas baseadas em preditores únicos ou múltiplos, assim como a modelagem de todos os semivariogramas necessários direto e cruzado.

Um outro ponto crucial que complementa os argumentos acima e também ratifica o termo sugerido é que os fenômenos devem ser observáveis em diferentes aspectos, tais como o padrão espacial pode variar quando se varia a direção, alguns fatores podem afetar o comportamento da variável e gerar uma tendência na região (MÜLLER, 2007), podem existir variáveis cujos comportamentos não atendem as hipóteses fundamentais da Geoestatística, etc (ISAAKS; SRIVASTAVA, 1989; SOARES, 2006). Finalmente, devido a todas essas importantes questões abordadas logo acima, considera-se o termo complexo espacialmente estocástico (cee) mais adequado que o uso do termo processo estocástico.

Essa coleção de variáveis aleatórias, legitimamente, representa a verdadeira população estocástica, a qual é o escopo central dos estudos da Estatística como um todo. Consideram-se como uma população estocástica todos fenômenos latentes disponíveis dentro da região $\mathcal{R}$. Essa ideia pode ser melhor compreendida analisando-se a Figura 2.2. Observe que o gráfico mostra uma infinidade de fenômenos latentes ocorrendo na região $\mathcal{R}$, sendo os mesmos representados pelas superfícies (ou conjunto de variáveis aleatórias) $\{Z_1(x), \ldots, Z_k(x), \ldots, Z_\alpha(x)\}$, indicando a existência de um processo multivariado composto de α fenômenos. Assim, cada superfície $Z_k(x)$ representa um único complexo espacialmente estocástico e o vetor espacial $x' = (x_1, x_2, \ldots, x_n)'$ sinaliza os n possíveis pontos amostrais tomados dentro da região $\mathcal{R}$. Percebe-se que cada componente do vetor espacial x' está implicitamente associado ao sistema de coordenadas do plano cartesiano que deve ser definido pelo pesquisador no planejamento amostral (WACKERNAGEL, 2003; SCHABENBERGER, 2005). Portanto, quando se fixa um ponto dentro de $\mathcal{R}$, tem-se uma única realização multivariada envolvendo k fenômenos de interesse e, dessa forma, obtém-se uma amostra multivariada de tamanho 1. Esse sistema de amostragem multivariado também está exemplicado na Figura

2.2, por meio de uma haste vertical branca, que ao seccionar as superfícies, gera-se um vetor amostral multivariado composto pelos seguintes valores das variáveis: $(z_1(x_i), \ldots, z_\alpha(x_i))'$, tal que $i = 1, \ldots, n$ e x_i define a particular coordenada de um ponto amostrado dentro do campo aleatório $\mathcal{R}$. Obviamente, realizando esse procedimento repetidamente, pode-se realizar qualquer tamanho de amostra.

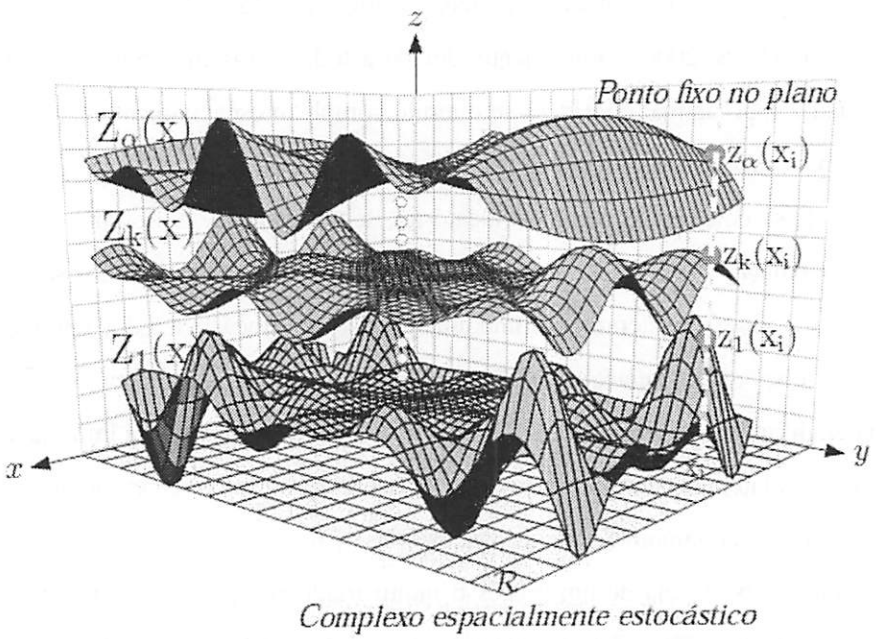


Figura 2.2 Complexo espacialmente estocástico: simulação de processos latentes ocorrendo conjuntamente em uma região

De forma geral, uma base de dados coletados a partir de um complexo espacialmente estocástico multivariado pode ser expressada e tratada matematicamente por meio da matriz de dados amostrais, $\mathbf{Z}_{n\times\alpha}$, formada pelos $(j,i)-$ésimos

elementos $z_j(x_i)$, conforme mostrado em (2.1):

$$\mathbf{Z} = \begin{bmatrix} z_1(x_1) & z_2(x_1) & \dots & z_k(x_1) & \dots & z_\alpha(x_1) \\ z_1(x_2) & z_2(x_2) & \dots & z_k(x_2) & \dots & z_\alpha(x_2) \\ z_1(x_3) & z_2(x_3) & \dots & z_k(x_3) & \dots & z_\alpha(x_3) \\ \vdots & \vdots & \vdots & \dots & \vdots & \\ z_1(x_n) & z_2(x_n) & \dots & z_k(x_n) & \dots & z_\alpha(x_n) \end{bmatrix}_{n \times \alpha} . \quad (2.1)$$

E, pode-se, ainda, reescrever a matriz em (2.1) de forma mais simplificada,

$$\mathbf{Z} = (z_1(x), z_2(x), \dots, z_\alpha(x))'.$$

Assim, esse modelo matricial composto de uma coleção de variáveis aleatórias contém todo o suporte estatístico para dar início aos procedimentos de predição geoestatística e, consequentemente, conhecer a verdadeira população estocástica, pertencente ao domínio $\mathcal{R}$, por meio de técnicas de interpoladores lineares ótimos (ISAAKS; SRIVASTAVA, 1989; CRESSIE, 1993; WEBSTER; OLIVER, 2001).

Conclusivamente, entende-se que esses fenômenos são gerados por algum mecanismo aleatório, mas com certa dependência espacial, o que dá a esses fenômenos o caráter de serem regionalizados. Tendo em vista essas propriedades intrínsecas a esse tipo de fenômenos, será necessário postular um modelo baseado na teoria das probabilidades para modelar esse conjunto de variáveis aleatórias. Assim, por causa da natureza dos fenômenos, não se pode analisar simplesmente o conjunto de dados realizados sem levar em conta a sua característica aleatória. Portanto, faz-se necessário a elaboração da teoria estocástica que está habilitada para modelar os fenômenos dessa natureza.

2.1.2 Formalização do modelo matemático

Nesse ponto deve estar claro que, na matriz de dados espaciais multivariada, existe correlação tanto nas linhas (entre os unidades ou pontos amostrais) quanto também entre as colunas. Assim, os processos (variáveis estocásticas) abordados no estudo diferem do que ocorre com a matriz de dados construída sob o suposto de independente e identicamente distribuídos, que teria possivelmente correlação só entre as colunas.

Toda a teoria Geoestatística está fundamentada no conceito de variável regionalizada. Então, os fenômenos tipicamente estudados ou de interesse possuem a principal característica de serem razoavelmente estruturados no espaço. Assim, a Geoestatística é um campo da Estatística Espacial que se preocupa em modelar processos estocásticos cujas variáveis aleatórias se distribuem continuamente por toda a região $\mathcal{R}$ do inquérito. O formato básico dos dados da Geoestatística univariada é $z(x_i) : i = 1,...,n$, em que x_i identifica a localização espacial e $z(x_i)$ representa a realização do processo (valor escalar ou variável resposta) associada ao local x_i.

Para uma formalização mais teórica das variáveis regionalizadas, a simbologia

$$z(x)' = (z(x_1),z(x_2),\ldots,z(x_n))'$$

para todo $x_i \in \mathcal{R}$, representa as n observações tomadas nos locais $\{x_1,\ldots,x_n\}$.

Por meio dos métodos da Geoestatística é possível solucionar uma variedade de problemas e a ponte que os unem deve-se ao fato de poderem ser pensados como uma realização de um processo estocástico. Mais formalmente, em aplicações Geoestatísticas, os dados são assumidos ser uma realização parcial de um

processo estocástico (coleção de variáveis aleatórias)

$$\{Z(x) : x \in \mathcal{R} \subset \Re^p\}$$

definido sobre um mesmo espaço de probabilidade, cujas variáveis aleatórias são indexadas em um subconjunto $\mathcal{R}$ do espaço vetorial p-dimencional $\Re^p$ (WALLER; GOTWAY, 2004). Este último é chamado espaço de índices do processo estocástico. O espaço de índices $\Re^p$ é definido de tal maneira que seja possível representar variações aleatórias em espaços de qualquer dimensão, por exemplo, $p = 1$ para variações no tempo (como os estudos de séries de tempo), $p = 2$ para variações em seperfícies, $p = 3$ para variações no espaço tridimencional e $p = 4$ para variação no espaço-tempo. De acordo com Waller e Gotway (2004, p. 273), para uma localização fixa x, $Z(x)$ é uma variável aleatória para qual as leis de probabilidades se aplicam; para uma realização fixa desse processo, observa-se uma função de espaço a saber, os dados nos locais $x_1, \ldots, x_n$. Os dados são somente uma realização parcial de uma função espacial visto que não se pode, por razões práticas, observar o processo em todos os pontos em $\mathcal{R}$.

A realização de um processo estocástico univariado, formado por uma coleção de variáveis regionalizada, tomadas dentro de uma área $\mathcal{R}$, está ilustrada na Figura 2.3.

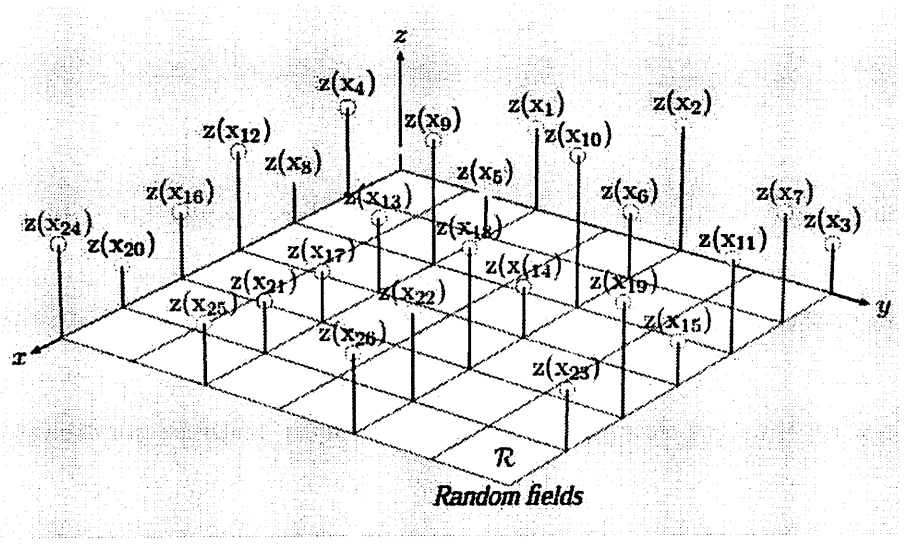


Figura 2.3 Simulação de um campo aleatório univariado

2.2 Geoestatística

> First law of geography:
> "Everything is related to everything else, but near things are more related than distant things."
>
> (TOBLER, 1969, p. 7)

Nesta seção será apresentada, primeiramente, uma contextualização histórica da Geoestatística, relatando seu nascimento e seu avanço (principalmente computacional) até os dias atuais. Além disso, serão apresentados os tópicos mais importantes relacionados com a fase da modelagem do padrão espacial. A predição espacial, portanto, não faz parte da pesquisa desta tese.

2.2.1 Contextualização e conceitos

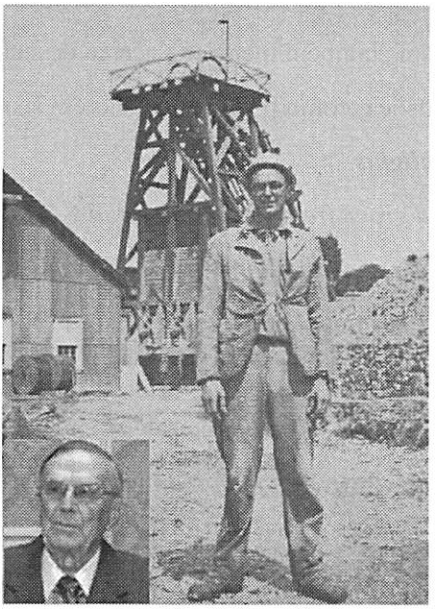


Figura 2.4 Daniel Krige: África do Sul, 1939

Nada é mais natural do que iniciar essa seção sobre Geoestatística comentando um pouco sobre o seu precursor, Daniel Gerhardus Krige, um sul-africano que, tendo graduado-se em Engenharia de Minas, em 1938, com apenas 19 anos de idade, começou o aprendizado de seu ofício a partir do ano de 1939 (Figura 2.4). No intuito de enfatizar sua grande importância no nascimento da Geoestatística, por incitar o também engenheiro de minas francês George Matheron com suas inovadoras técnicas para inferir e mapear recursos minerais, destaca-se a seguir a participação de Daniel G. Krige em uma das seções plenárias do 33º International Symposium on Application of Computers and Operations Research in the Mineral Industry (33º APCOM), realizado no Departamento de Engenharia de Minas da Universidade do Chile, em Santiago. Esse simpósio ocorreu no período de 24 a 27 de abril de 2007, sendo sua primeira vez na América Latina. Nesse referido evento (seção plenária de Geoestatística), Daniel G. Krige apresentou uma revisão prática de alguns conceitos básicos e técnicas para aplicações em minas, na qual ele relata, resumidamente, suas experiências, que foram a principal inspiração que desencadeou os estudos e posterior formalização teórica que gerou os consagrados métodos da Geoestatística desenvolvidos por Georges Matheron. Assim, traz-se uma parte original

desses escritos apresentados por Daniel G. Krige, *The origins and development of Geostatistics* (KRIGE, 2007), no qual ele expressou sua gradidão pelas oportunidades, como segundo ele mesmo afirma, que foram dádivas concebidas por meio da infinita graça que vieram do ALTO. O mesmo também menciona acerca de sua colaboração no desenvolvimento da Geoestatística, como pode ser notado em seu pequeno relato abaixo: ***Grateful for the highlights***

–Opportunities, all flowing through unlimited grace from ABOVE, to study, to do research, to publish papers, to meet and co-operate with international colleagues, good health throughout and constant support from my family, two wonderful wives, and many colleagues.

–My career in Geostatistics started in 1950, almost 60 years ago - was involved in the birth of Geostatistics and its establishment worldwide in mining and other fields.

–Geostatistics is still growing and reminds us of Socrates, probably the most knowledgable philosopher of his time who said: "I know nothing, all I really know is that I am ignorant". Obviously, somewhat of an overstatement but in our situation it can be accepted as indicating a vision of unlimited scope for further knowledge via useful research in this field, new developments and the proper application of existing techniques.

–1951/2: De Wijs (Netherlands) — study of differences between values and different lags — idea of Variogram.

–1955/60's: French translation of these two papers (1955), plus contributions by Matheron and Duval, a paper by Allais on exploration prospects in the Sahara (1956) based on the lognormal model, start of French School of Geostatistics in Paris(Allais, Matheron); Matheron's work and publications on the theory of regionalised variables

–1964: 4th APCOM, Colorado School of Mines. Early equivalent of computerised Kriging of ore reserve blocks and routine computer applications on Anglovaal Group mines.

–1965/66: Matheron's insistence on using the term Kriging.

–1969: 8th APCOM, Salt Lake City; attended by author and Matheron — Kriging accepted partly, but not yet fully in U.S.A. – contact with Rendu.

–Macro (elephant) kriging: the reference paper on macro — kriging by myself and Dr Assibey-Bonsu: APCOM 1992, Tucson, Arizona. My colleague presented the paper and gave the name "Elephant" Kriging to the technique.

Figura 2.5 George Matheron, 1988

A Geoestatística teve suas origens no início dos anos cinquenta, a partir de inquéritos sobre a elaboração de métodos de estimação da concentração de ouro que considerasse a existência da variabilidade espacial desse particular fenômeno de interesse. Esses estudos foram conduzidos devido às necessidades práticas vividas pelas indústrias de mineração da África do Sul. Os primeiros estudos foram realizados pelo engenheiro de minas Daniel G. Krige e o estatístico H. S. Sichel (KRIGE, 1951). Essas técnicas chamaram a atenção dos engenheiros de minas do Centre de Morphologie Mathématique in Fontainebleau, França e, em especial, a de George Matheron, que aprimorou os conceitos inovativos de Krige (procedimentos de médias móveis para dados espacialmente correlacionados), combinando-os aos conhecimentos básicos até então adquiridos da Estatística convencional. A partir desse suporte teórico, Matheron desenvolveu uma única e poderosa estrutura de

análise espacial denominada Teoria das Variáveis Regionalizadas (MATHERON , 1971). Essa propriedade de o fenômeno ser regionalizado confere aos dados a qualidade de se distribuírem de forma estruturada, a princípio continuamente, sobre toda a região $\mathcal{R}$ do inquérito, isto é, os dados, possivelmente, possuem dependência espacial e/ou temporal. Formulando com outras palavras: os locais vizinhos tendem ser parecidos, não atípicos. Essa nova metodologia passa, então, a dar a devida importância à questão da variabilidade espacial que pode estar ocorrendo subjacente ao inquérito de estudo. Perceba que a variabilidade espacial não pode ser detectada pelos métodos da Estatística convencional, o que a torna inviável para resolver problemas de quaisquer áreas da ciência que necessitam modelar fenômenos que possuem padrões de dependência espacial. Certamente, nesse contexto, a Geoestatística surge como uma alternativa de análise suportada por um poderoso conjunto de técnicas que podem ser sintetizadas por meio de duas fases, a saber:

Fase da caracterização: consiste da parte mais importante da análise Geoestatística. Nessa primeira fase, investigando-se adequadamente a amostra de dados georreferenciados, pode-se verificar a existência da dependência espacial (ou independência espacial), bem como identificar o tipo de padrão espacial que está ocorrendo na região, $\mathcal{R}$, por meio do gráfico denominado variograma ou semivariograma experimental (também conhecido como semivariograma amostral). Em seguida, busca-se modelar o fenômeno ajustando-se, adequadamente, um dos modelos teóricos usuais autorizados ao gráfico do semivariograma experimental, com o qual se estimam os parâmetros de covariância, a saber: efeito pepita, patamar e alcance. De modo mais geral, esta fase é também conhecida como análise descritiva do fenômeno.

Fase da predição: é verdadeiramente considerada o coração da Geoestatística por

ser o principal objetivo da maioria dos estudos em diversas áreas e, ainda, pelo interesse de cunho prático. Consiste de uma extensão de métodos conhecidos por krigagem ("kriging", em inglês), com os quais se podem predizer valores pontuais, médias de blocos (subregiões) ou, ainda, interpolar toda a região de estudo, considerando um conjunto finito de valores observados em locais espaciais (coordenadas) sobre padrões de amostragem regular ou irregular. Toda a predição é realizada mediante interpolação linear (método de krigagem), a qual possui propriedades ótimas que lhe conferem o título de melhor preditor linear não viesado e de variância mínima (*best linear unbiased preditor*; BLUP).

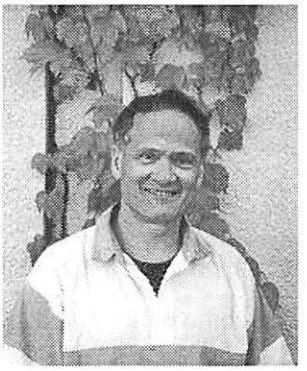


Figura 2.6 Andre Journel

A Geoestatística foi "popularizada" nos Estados Unidos por Andre G. Journel, considerado um dos maiores geoestatísticos em todo o mundo. Atualmente ele é professor da Escola de Ciência da Terra, na Universidade de *Stanford*, Califórnia, EUA, onde é conhecido por *Donald and Donald M. Steel Professor*. Journel concluiu *PhD* em matemática aplicada, mas sua consagração profissional veio por meio da Geoestatística, de modo que se tornou o principal responsável por uma abordagem inovadora, a qual ele mesmo diz ser uma reedição para uma Geografia quantitativa e preditiva. Segundo Journel, a Geografia era muito descritiva: "Geógrafos observaram as formas da Terra e, então, disseram que elas foram geradas pelas Eras Glaciais ou erosão ou seja lá o que for". Porém, as observações permitem fazer predição e a visão anterior era praticamente descritiva. Assim, esse conceito pode ser ratificado por uma importante afirmação de Jounel que revolucionou a

forma de se pensar a Geografia, trazendo-a radicalmente para o universo da Estatística. O texto original é transcrito a seguir:

> *Geographers missed the quantitative and computer revolution. It's not enough to make a nice picture. You must translate that into numbers so you can pass it on to the engineers. Engineers can't work with pictures; they need numbers.*

Realmente, os trabalhos de Journel foram de grande importância no processo de consolidação da Geoestatística como uma ciência digna de respeito. Uma prova disso foi o seu livro *Mining Geostatistics*, publicado em 1978, que foi o primeiro trabalho completo de Geoestatística voltada para a engenharia de minas. Esse livro sintetiza as experiências adquiridas pelos pesquisadores do Centro de Morfologia e Matemática na França, além de engenheiros de minas e geologistas de todo o mundo que deram sua contribuição (JOURNEL; HUIJGBREGTS, 1991).

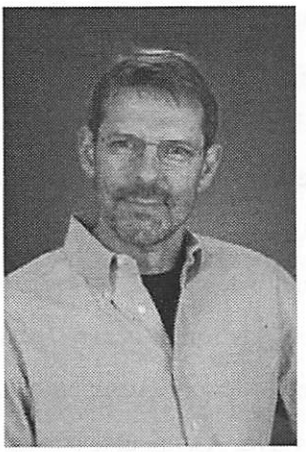


Figura 2.7 Noel Cressie

Um passo à frente importante para a aceitação plena da Geoestatística como uma parte da Estatística foi o trabalho de Noel A. C. Cressie, no ano de 1993, em seu conceituado livro, visto como referência básica (CRESSIE, 1993), no qual disponibilizou aos cientistas e engenheiros razoáveis métodos estatísticos para análise de dados espaciais. Sendo estatístico, Noel Cressie contribuiu grandemente para o desenvolvimento da Estatística espacial como um todo, construindo uma abordagem genuinamente estatística, algo que até aquela época

não se havia feito ainda. Pode-se dizer que, com esse livro, Cressie abriu novos horizontes no contexto da Estatística Espacial, apresentando uma abordagem que enveredou pelos já consolidados métodos da Estatística convencional, dando ênfase a uma linguagem mais adequada ao universo estatístico. Suas ideias foram inovadoras e projetaram uma linguagem mais condizente com o universo estatístico que, em linhas gerais, originou uma clara tendência de o pensamento científico reunir esforços para dar, em particular, à Geoestatística, um caráter propriamente estatístico no processo de modelagem espacial.

Assim, a Geoestatística passa, então, a gozar das prerrogativas da estatística convencional, tais como: modelos baseados em verossimilhança, modelos lineares generalizados, normalidade multivariada, testes de hipótese, modelos mistos generalizados, abordagens bayesianas, etc. (DIGGLE; RIBEIRO JUNIOR, 2007). Essa atitude marcou o início de um desenvolvimento teórico que culminou em um grande avanço nos procedimentos inferenciais da Geoestatística, buscando modelos estocásticos mais explícitos, os quais são centrais para a estatística moderna e, com isso, a Geoestatística está ampliando cada vez mais o seu leque de aplicação e, conseqüentemente, tornando-se bastante poderosa e útil para diversas áreas do desenvolvimento científico. Esta tese caminha nessa mesma direção, de dotar a Geoestatística dos mesmos procedimentos e ideias que já mostram seu valor na Estatística convencional.

Uma das importantes contribuições de Noel Cressie que merecem destaque é a classificação da Estatística Espacial em três principais classes de modelos, a saber: padrão (espacial) de pontos, dados *lattice* ou dados de área (variações espaciais discretas) e a Geoestatística. Em especial, problemas típicos de inferência Geoestatística se diferenciam dos demais (Dados *lattice*, Padrões de pontos) pelo fato de o índice espacial variar continuamente sobre a região $\mathcal{R}$ do inquérito (no

jargão de alguns: a área de estudo deve estar "contaminada" pelo fenômeno de interesse). Mas, essa particular propriedade dos modelos da Geoestatística não os impossibilita de serem adotados a partir de problemas usualmente abordados em qualquer outra classe de modelos. Nesse aspecto, a Geoestatística ainda precisa ser mais explorada e, com certeza, há um vasto e promissor campo de pesquisa.

De fato, o prefixo "geo", em Geoestatística, denuncia a origem de seus métodos, os quais foram elaborados especificamente para solucionar problemas relacionados à ciência da Terra, isto é, a prática geoestatística é *ad hoc* por natureza. Porém, os métodos da Geoestatística constituem um poderoso conjunto de técnicas com um enorme potencial que podem auxiliar pesquisas em vários ramos da ciência. Por conseguinte, a Geoestatística necessita romper o cordão umbilical original e ampliar seus conceitos concernentes às teorias estatísticas mais universais, bem como buscar maior abrangência em suas aplicações para índices espaciais contínuos (CRESSIE, 1993).

Avançando mais no aperfeiçoamento da modelagem geoestatística, Diggle, Tawn e Moyeed (1998) foram os primeiros a cunhar a expressão *Model-based Geostatistics* (Geoestatística baseada em modelo). Como pioneiro, Peter Diggle escreveu vários trabalhos discutindo esse tema, em que se destacam os seguintes trabalhos: Diggle et al. (2002); Diggle, Ribeiro Junior e Christensen (2003) e, mais recentemente publicou, juntamente com Paulo Justiniano Ribeiro Junior, o livro *Model-based Geostatistics* (DIGGLE; RIBEIRO JUNIOR, 2007).

Realmente, a Geoestatística baseada em modelos já se estabeleceu como uma abordagem promissora e vem ganhando interesse no contexto da análise de dados espacialmente contínuos. O seu poder de análise se justifica pela possibilidade de procedimentos inferenciais que já são consagrados na Estatística convencional, mas que, agora, estão modestamente sendo aplicados para solucionar

uma gama de problemas cujos fenômenos possuem razoável dependência espacial. Construindo-se o processo estocástico regionalizado sobre uma estrutura realmente probabilística, em que a variável aleatória possa ser modelada com base nos fundamentos da teoria das probabilidades, o procedimento inferencial sobre o comportamento espacial ganha um caráter mais propriamente estatístico. Com esse enfoque, modelos probabilísticos são adotados a priori e, a partir desses supostos modelos, métodos são adequados à buscar estimativas não viesadas e de variância mínima.

Contemporaneamente, as literaturas sobre Estatística estão cada vez mais inserindo modelos espaciais em seus textos. Já se podem observar trabalhos realizados em ciências agronômicas, engenharia, geologia, hidrologia, petróleo, ciência do solo, astronomia, economia, floresta, epidemiologia, etc. (veja os textos de Arbia (2006); Müller (2007)). Notadamente, as técnicas de Geoestatística têm potencial para serem aplicadas em qualquer área que trabalha com dados coletados em locais espaciais e que necessitam desenvolver modelos habilitados para detectar a existência (ou ausência) da dependência espacial entre medições em diferentes localizações x_i. Porém, infelizmente, a Geoestatística ainda carrega o peso de suas bases originais, sendo tratada por muitos pesquisadores como limitada e que seus métodos devem ser adotados apenas em casos bem específicos da Estatística espacial.

2.3 Estacionaridade do processo estocástico e tendência nos dados

Trazendo à tona a teoria das distribuições de probabilidades, a variável aleatória $Z(x_i)$ agindo em um ponto x_i da região $\mathcal{R}$ gera realizações seguindo

uma distribuição de probabilidades F, tal que

$$P(Z(x_i) \leq z) = F_{x_i}(z),$$

em que P é a probabilidade de uma realização de Z em um ponto x_i ser menor que um valor fixo z (WACKERNAGEL, 2003).

No caso bivariado, por exemplo, uma função de distribuição de probabilidades para duas variáveis aleatórias $Z(x_1)$ e $Z(x_2)$, tomadas em locais diferentes, é dada por

$$P(Z(x_1) \leq z_1, Z(x_2) \leq z_2) = F_{x_1,x_2}(z_1, z_2),$$

em que P é a probabilidade simultânea do resultado de $Z(x_1)$ ser menor que z_1 e um resultado de $Z(x_2)$ ser menor que z_2.

Generalizando, uma função de distribuição conjunta (ou múltipla) para n variáveis aleatórias localizadas em n diferentes locais pode ser definida como

$$F_{x_1,\ldots,x_n}(z_1, \ldots, z_n) = P(Z(x_1) \leq z_1, \ldots, Z(x_n) \leq z_n). \qquad (2.2)$$

De acordo com Arbia (2006), a expressão dada em (2.2) refere-se a uma das características fundamentais de um campo aleatório, que é a estrutura de dependência da variável aleatória $\{Z(x_i), x_i \in \mathcal{R}\}$; a outra característica fundamental de um campo aleatório diz respeito à natureza do índice x_i, isto é, está relacionada à topologia das observações (caso seja necessário, consultar Arbia (2006, p. 36), para obter detalhes sobre esse assunto). Segundo Wackernagel (2003), a teoria construída anteriormente é um extraordinário modelo geral que é capaz de descrever qualquer processo real ou tecnológico. Porém, na prática, têm-se em poder somente poucos dados a partir de uma ou várias realizações da variável ale-

atória. Assim, é impossível conhecer a distribuição conjunta e, consequentemente, impossível inferir sobre todas as funções de distribuição uni ou multivariada de qualquer conjunto de dados. Por isso, a simplificação é necessária e a saída para o problema está no conceito de estacionaridade.

No contexto da Geoestatística, diz-se que uma variável aleatória (regionalizada) $Z(x_i)$ é estritamente estacionária se, para qualquer conjunto de n pontos $x_1, \ldots, x_n$ pertencente a um campo aleatório e para qualquer distância h,

$$F_{x_1,\ldots,x_n}(z_1,\ldots,z_n) = F_{x_1+h,\ldots,x_n+h}(z_1,\ldots,z_n). \tag{2.3}$$

Dessa forma, estacionaridade significa que uma translação conjunta de uma configuração de pontos não muda a distribuição conjunta do processo subjacente. Esse fato é denominado de translação invariante. Arbia (2006, p. 44) interpreta (2.3) da seguinte forma: se um campo aleatório permanece inalterado, em termos de sua função de densidade de probabilidade conjunta depois de uma translação, diz-se que ele é estacionário sob translação, ou homogêneo, o que implica que o tipo de estrutura de dependência dentro do campo aleatório não muda sistematicamente de um lugar para outro. Na verdade, considerar que um fenômeno é estritamente estacionário é uma suposição bastante forte e remota, tendo em vista assumir que todos os parâmetros da função aleatória são invariantes de ponto a ponto na área e, certamente, é um tremendo desafio descrevê-los.

Naturalmente, sempre há um certo grau de incerteza de magnitude sobre todas as coisas (fenômenos ou processos físicos) no universo e a prova disso é a própria ciência Estatística que surgiu para tentar minimizar essas incertezas sobre as gandrezas universais. A dúvida sempre assombrou o homem (WACKERNAGEL, 2003). Em vista disso, deve-se relaxar o conceito de estacionaridade estrita e definir vários tipos e graus de estacionaridade dos processos ocultos gerando as

variáveis regionalizadas. Obviamente, o tipo de estacionaridade assumida determina o tipo de inferência estatística ou krigagem que é permitida. Somente para ratificar, Arbia (2006, p. 45) declara que esse tipo de estacionaridade, porém, é muito raramente realizada em circunstâncias empíricas e, por essa razão, tem-se que introduzir o conceito mais fraco de estacionaridade de ordem k que pode ser definida sobre as bases dos k primeiros momentos do campo aleatório.

É importante entender que estacionaridade é uma propriedade do modelo da função aleatória (modelo probabilístico) e não da variável regionalizada. Wackernagel (2003) resalta que, na prática, pode-se dizer que uma dada "variável regionalizada é estacionária", mas isso é claro, sempre significando uma taquigrafia para: esta variável regionalizada pode ser considerada uma realização de um modelo de uma função aleatória estacionária.

2.3.1 Estacionaridade dos dois primeiros momentos

Como visto na secão 2.3, a estacionaridade estrita é raramente assumida porque requer a descrição da distribuição conjunta (2.3) para qualquer conjunto de pontos $\{x_1, \ldots, x_n\}$, o que é um dificultador. Então, a formidável alternativa será considerar somente os pares de pontos $\{x_1,x_2\}$ para determinar apenas os dois primeiros momentos e não toda a distribuição. A seguir serão definidas as duas hipóteses básicas exigidas nos modelos de Geoestatística (WEBSTER; OLIVER, 2001; SCHABENBERGER; GOTWAY, 2005; WACKERNAGEL, 2003):

estacionaridade de segunda ordem: essa hípotese é assumida para os dois primeiros momentos da variável. A estacionaridade de segunda ordem existe se a média e a variância da função aleatória são independentes da localização e a covariância depende somente da distância ou do incrememto entre

os valores medidos da variável regionalizada.

estacionaridade intrínseca: trata-se de uma suposição mais fraca e assume-se apenas a estacionaridade dos dois primeiros momentos da diferença de um par de valores em dois pontos e conduz ao conceito do semivariograma.

Mais precisamente, conforme definido por Webster e Oliver (2001), estacionaridade significa que a distribuição dos processos aleatórios tem certos atributos que são os mesmos em qualquer local. Sendo assim, dado o processo $\{Z(x) : x \in \mathcal{R} \subset \Re^p\}$, para o primeiro momento, assume-se que a média, $\mu(x) = E[Z(x)]$, sobre a qual as realizações individuais oscilam, é constante para todo x. Essa construção permite substituir cada $\mu(x), \forall\ x$, por um único valor μ, o qual pode ser estimado por amostragem repetitiva. Desse modo, cada realização no ponto x é considerada uma repetição do experimento. Note que, ao admitir que todas as variáveis aleatórias têm a mesma média μ, esse parâmetro passa a ser não dependente da posição x e pode ser estimado por meio da média aritmética dos valores das realizações das variáveis aleatórias (SOARES, 2006):

$$\hat{\mu} = m = \frac{1}{n}\sum_{i=1}^{n} z(x_i).$$

Soares (2006) menciona que essa estacionaridade da média não pode nunca ser validada ou refutada, uma vez que, na realidade, só existe uma realização da função aleatória. Esse autor, contudo, afirma que ela pode ser julgada apropriada, ou não, dependendo da homogeneidade da amostra na área $\mathcal{R}$ em que a variável se distribui e que, de fato, a hipótese de estacionaridade da média implica que esta pode ser estimada pela média aritmética. Ainda, julgar esta hipótese de estacionaridade como apropriada é julgar a média das amostras como representativa da mesma região $\mathcal{R}$. Em outras palavras, é considerar que os valores das amostras

são suficientemente homogêneos para validar aquela representatividade.

É sabido que o semivariograma, que será definido posteriormente, ou covariância são medidas da correlação espacial entre duas variáveis aleatórias. Para o 2° momento, a hipótese de estacionaridade de 2^a ordem dessas estatísticas, considerando quaisquer duas variáveis $Z(x_i)$ e $Z(x_j)$, separadas por um vetor de distância h, implica:

$$\begin{aligned} Cov[Z(x_i+h),Z(x_i)] &= E[\{Z(x_i+h)-\mu\}\{Z(x_i)-\mu\}] \\ &= E[\{Z(x_i+h)\}\{Z(x_i)\}-\mu^2] \\ &= C(h). \end{aligned} \tag{2.4}$$

Schabenberger e Gotway (2005) afirmou que a existência da função de covariância $C(h)$ em um campo aleatório estacionário de segunda ordem tem importantes consequências. Visto que $C(h)$ não depende das coordenadas e

$$Cov[Z(x_i),Z(x_i+0)] = Var[Z(x_i)] = C(0),$$

seque que a variabilidade de um campo aleatório estacionário de segunda ordem é a mesma em todo parte de $\mathcal{R}$.

De acordo com Webster e Oliver (2001), se o processo espacial, $\{Z(x) : x \in \mathcal{R} \subset \Re^p\}$, tem uma distribuição gaussiana (uma distribuição normal multivariada) para todos os pontos do campo aleatório $\mathcal{R}$, então, a média e a função de covariância caracterizam completamente o processo.

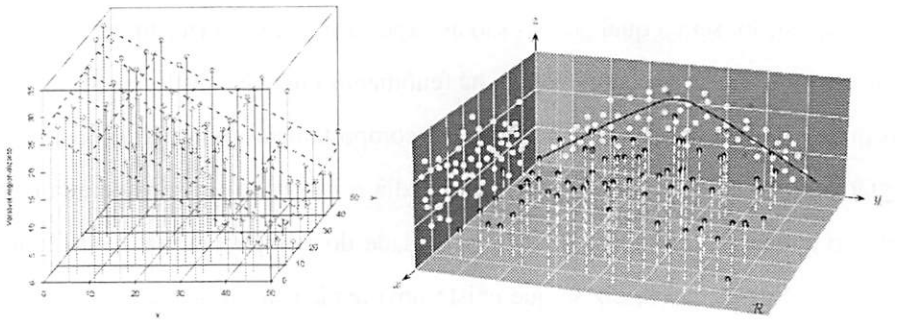
Wackernagel (2003) define estacionaridade como sendo a propriedade que as variáveis aleatórias possuem de se manterem as mesmas, quando se muda de posição um dado conjunto de n pontos, de um local para outro da região $\mathcal{R}$, onde

ocorre o processo de estudo.

Sabe-se que a estacionaridade de um processo estocástico é uma importante suposição, sem a qual poucas são as expectativas ao se fazer inferência, com base em uma amostra. Naturalmente, há fenômenos em que os valores dos atributos que se pretendem predizer não têm um comportamento homogêneo dentro da região amostrada. Em outras palavras, a média μ não pode ser considerada constante, o que anula a hipótese de estacionaridade do primeiro momento da função aleatória. Nesses casos, diz-se que existe uma tendência ou deriva ocorrendo no processo estocástico $\{Z(x) : x \in \mathcal{R} \subset \Re^p\}$ (SOARES, 2006). Têm-se como exemplos:

- os dados possuem algum tipo de gradiente em uma dada direção;
- notam-se altas concentrações de valores localmente num ponto ou em um setor da região amostrada;
- valores crescendo ou diminuindo de modo sistemático em uma direção específica.

Na Figura 2.8 são apresentadas ilustrações de duas possíveis situações de deriva ocorrendo em um determinado processo espacial, caracterizando que a média não pode ser considerada constante na região de estudo, $\mathcal{R}$. Claramente, pode ser visto um caso em que o comportamento sistematicamente linear (Figura 2.8a) e um outro fenômeno apresentando um comportamento parabólico (Figura 2.8b). Observa-se na Figura 2.8a, que o plano clinal (gradiente) indica que média deve ser modelada por uma função polinomial do primeiro grau e o método mais comum para estimar uma deriva é o uso de ajuste de tendência por quadrados mínimos.



(a) Deriva linenar (b) Deriva quadrática

Figura 2.8 Gráficos exemplificando possíveis tipos de tendência ou deriva frequentemente presente em processos espaciais

2.3.2 Semivariograma

> "Statistics, the science of uncertainty, attempts to find structure in chaos."
>
> (CRESSIE, 1989)

De antemão, faz-se necessário compreender na íntegra como se dá a caracterização da dependência espacial por meio de instrumentos de medidas denominadas dissimilaridades/similaridades. Em outras palavras, nesta seção trata-se dos procedimentos de estimação do semivariograma ou variograma, o qual é o alicerce da Geoestatística.

2.3.2.1 Semivariogramas de nuvens: dissimilaridade versus distância

Para qualquer conjunto de variáveis aleatórias, $\{Z(x), x \in \mathcal{R}\}$, pode-se calcular a semivariância para cada par de pontos x individualmente como (WEBSTER; OLIVER, 2001)

$$\hat{\gamma}_{ij}(h) = \frac{1}{2}[z(x_i) - z(x_j)]^2. \tag{2.5}$$

Essa quantidade mede a dissimilaridade entre pares de valores de dados em um campo aleatório, efetuando o cálculo da metade do quadrado da diferença entre dois valores $z(x_i)$ e $z(x_j)$. Note que a natureza topológica de dois pontos x_i, x_j (geometria do espaço geográfico) permite ligá-los por um vetor $h = x_i - x_j$ (WACKERNAGEL, 2003). Consequentemente, tem-se que a dissimilaridade $\hat{\gamma}_{ij}(x_i,x_j)$ depende do espaço ou distância e da orientação dos pares de pontos descritos pelo vetor h,

$$\hat{\gamma}_i(h) = \frac{1}{2}[z(x_i + h) - z(x_i)]^2. \tag{2.6}$$

Esses valores podem ser representados graficamente em função do vetor de distâncias h (*lags*) como um diagrama de dispersão, denominado de semivariograma de nuvens (ou *cloud*). Na Figura 2.9 observa-se um semivariograma de nuvens didático, supondo vários *lags* (grande variabilidade em h) implicando em pequenos incrementos no eixo das abscissas — gráfico tipicamente obtido em *lattice* irregular. Segundo Webster e Oliver (2001), o semivariograma de nuvens contém toda a informação sobre o relacionamento espacial nos dados em um dado *lag* e que, a princípio, poder-se-ia ajustar um modelo a ele para representar o semivarigrama na região, mas, na prática, é quase impossível julgar, a partir dele, se há alguma correlação espacial presente, que forma ela pode ter, e como ela poderia

ser modelada.

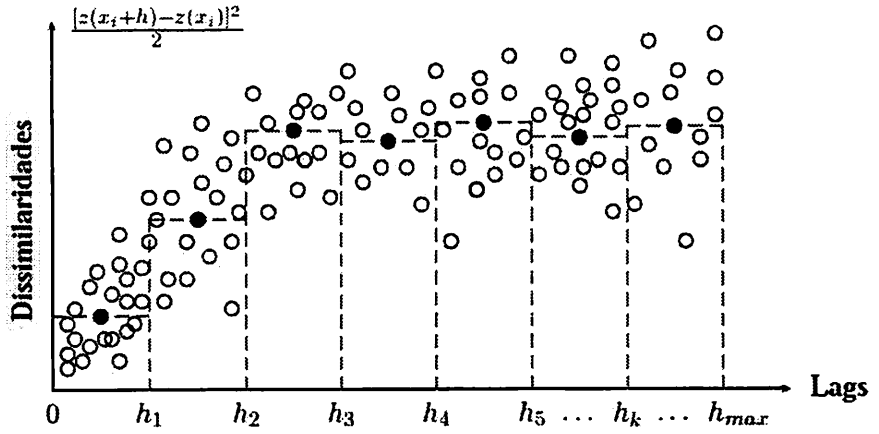


Figura 2.9 Semivariograma de nuvens mostrando as dissimilaridades por classe de distâncias

2.3.2.2 Semivariograma experimental: semivariância média

O estudo de modelos de correlação entre dados distribuídos espacialmente é denominado de análise estrutural ou modelagem do semivariograma, a ferramenta central da Geoestatística. O estimador clássico do semivariograma, proposto por Matheron (1962), é dado pela equação,

$$\hat{\gamma}(h) = \frac{\sum_{i=1}^{N(h)} \hat{\gamma}_i(h)}{2N(h)} = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i + h) - z(x_i)]^2, \qquad (2.7)$$

em que, $\hat{\gamma}(h)$ é donominado de semvariograma, $Z(x_i + h)$ e $Z(x_i)$ são variáveis regionalizadas, $N(h)$ determina o número de pares de valores medidos, das variáveis em estudo, separados por um vetor distância h. O gráfico plotado em relação aos correspondente valores de h é denominado semivariograma. Esse é um es-

timador não tendencioso, mas pouco resistente a observações atípicas devido ao termo ao quadrado no somatório. Cressie (1993, p. 75) apresentou um estimador mais robusto para o cálculo das variâncias, utilizando o ajuste de transformações de potência proposto por Box e Cox (1964), dado pela equação:

$$\bar{\gamma}(h) = \frac{1}{|2N(h)|} \sum_{N(h)} \{|z(x_i) - z(x_i + h)|^{\frac{1}{2}}\}^4 / 0{,}457 + \frac{0{,}494}{|2N(h)|}. \qquad (2.8)$$

Assim, o semivariogram é uma função do processo espacial que satisfaz às seguintes propriedades (SCHABENBERGER; GOTWAY, 2005):

- $\hat{\gamma}(-h) = \hat{\gamma}(h)$ [isto é, a autocorrelação entre $Z(x_i)$ e $Z(x_j)$ é a mesma que entre $Z(x_j)$ e $Z(x_i)$].
- $\hat{\gamma}(0) = 0$, visto que, por definição, $Var[Z(x) - Z(x)] = 0$.
- $\frac{\hat{\gamma}(h)}{\|h\|^2} \longrightarrow 0$ quando $\| h \| \longrightarrow \infty$, em que $\| h \|$ denota a dimensão do vetor h; isto é, $\hat{\gamma}(h)$ cresce mais lentamente que h^2.
- $\hat{\gamma}(\cdot)$ deve ser condicionalmente negativo definido, isto é,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \hat{\gamma}(x_i - x_j) \leq 0$$

para qualquer número finito de locais $\{x_i : i = 1, \ldots, n\}$ e números reais satisfazendo $\sum_{i=1}^{n} \lambda_i = 0$. Essa condição é semelhante à condição definida positiva para a função de covariância e matriz de variância-covariância, assegurando que todas as variâncias seja não-negativas.

A aplicação do variograma como medida de continuidade espacial requer apenas que os dados satisfaçam à hipótese intrínseca para uma variável regiona-

lizada (JOURNEL; HUIJBREGTS, 1991). Processos estacionários de segunda ordem são também intrinsecamente estacionários, isto é, a variância nos dados deve ser somente uma função da distância entre as amostras. Assim, a hipótese intrínseca pode ser matematicamente formulada como:

$$E[Z(x)] = m, \ \forall \ x \in \mathcal{R}$$

$$E[Z(x+h) - Z(x)] = 0$$

$$Var[Z(x+h) - Z(x)] = 2\gamma(h).$$

Pode ser demonstrado que a estacionaridade de segunda ordem implica na hipótese intrínseca, mas o inverso não é verdadeiro. No caso da estacionaridade de segunda ordem, tem-se (CRESSIE, 1993):

$$Var[(Z(x_i+h) - Z(x_i))^2] = E\{[(Z(x_i+h) - m) - (Z(x_i) - m)]^2\} =$$

$$= Var\{Z(x_i+h)\} + Var\{Z(x_i)\} - 2\{[Z(x_i+h) - m][z(x_i) - m]\} =$$

$$= 2C(0) - 2C(h) = 2\hat{\gamma}(h) \quad \text{ou}$$

$$\gamma(h) = C(0) - C(h).$$

Conforme Schabenberger e Gotway (2005), por causa da realação $\hat{\gamma}(h) = C(0) - C(h)$, os métodos estatísticos para campos aleatórios estacionários de segunda ordem podem ser delineados em termos da função de covariância ou do semivariograma. Enquanto os estatísticos são mais familiares com as variâncias e covariâncias, muitos geoestatísticos preferem o semivariograma.

As vezes, é de interesse particular comparar dois processos espaciais. Essa comparação é possível quando se utiliza a medida de correlação em vez da covari-

ância. Portanto, é bastante útil definir o correlograma espacial, como

$$\rho(h) = \frac{C(h)}{C(0)}.$$

Obviamente, o princípio de ρ é similar ao da típica correlação, isto é, o valor (ou quantidade) é escalado tal que $-1 \leq \rho \geq 1$ (WALLER; GOTWAY, 2004).

Um semivariograma típico com características ideais, está ilustrado na Figura 2.10. O seu padrão representa o que, intuitivamente, se espera de observações no campo. Isto é, espera-se que a diferença $Z(x_i + h) - Z(x_i)$ decresça à medida que h, a distância que os separa decresce. Isto é, espera-se que observações mais próximas geograficamente tenham um comportamento mais semelhante entre si que aquelas separadas por distâncias maiores. Note, ainda, que, por meio da Figura 2.10, podem ser identificados os principais parâmetros do semivariograma definidos a seguir.

Alcance: o parâmetro alcance (φ) é compreendido como sendo a distância dentro da qual as amostras apresentam-se correlacionadas espacialmente. Matematicamente, o valor do alcance é a abscissa do eixo de distâncias h correspondente ao valor da ordenada em que o gráfico, inicialmente, atinge a invariância, isto é, em que o variograma atinge o seu patamar. Deste ponto em diante considera-se que não existe mais dependência espacial entre os pontos amostrados porque a variância da diferença entre os pares de amostras $Var[Z(x_i + h) - Z(x_i)]$ torna-se invariante com a distância.

Patamar: denotado por $\sigma^2 = (\tau^2 + \varrho^2)$, o patamar consiste no valor da semivariância correspondente ao parâmetro alcance. Sob covariância estacionária, o patamar é equivalente à variância da variável aleatória regionalizada Z.

Efeito pepita: esse termo tem sua gênese e generalização a partir de seu signifi-

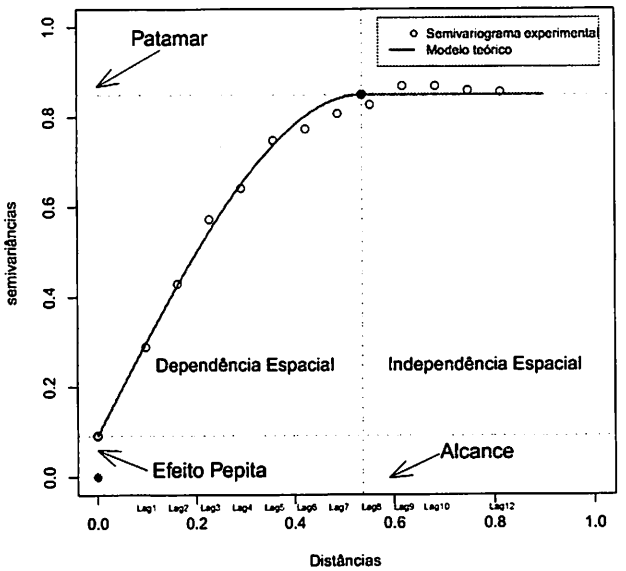


Figura 2.10 Semivariograma experimental com modelo teórico ajustado: o gráfico representa um semivariograma ideal, $\gamma(h)$, com as semivariâncias se estabilizando com o aumento do vetor distância h

cado em depósitos de minas de ouro. Idealmente igual a zero, entretanto, na prática, à medida que h tende a 0 (zero), se aproxima de um valor positivo chamado efeito pepita (τ^2), que revela sua aparente descontinuidade do variograma (ou covariograma) para distâncias menores que a menor distância entre as amostras.

Contribuição: representada por (ϱ^2), é a diferença entre o patamar e o efeito pepita. É o segmento gráfico nesse intervalo que, praticamente, caracteriza a dependência espacial de processos estocásticos contínuos.

A partir da Geoestatística, a ferramenta de segundo momento que caracteriza o processo estocástico espacialmente contínuo é o semivariograma. Assim,

o sumário padrão dos fenômenos regionalizados $Z(x)$ pode ser realizado por sua função de covariância $C(x_i,x_j) = \text{Cov}\{Z(x_i) - Z(x_j)\}$ ou por sua função de semivariograma $\gamma(x_i,x_j) = \frac{1}{2}Var\{Z(x_i) - Z(x_j)\}$.

2.4 Método de ajuste do semivariograma

Sabe-se que em estudos da caracterização do padrão de dependência espacial, o estimador do semivariograma experimental, $\hat{\gamma}(h)$, é uma função empírica essencial. Então, por meio do semivariograma, pode-se explorar o comportamento do fenômeno espacial subjacente existente em uma área $\mathcal{R}$, bem como realizar estimativas dos parâmetros de dependência espacial $\theta = (\hat{\tau}^2, \hat{\sigma}^2, \hat{\varphi})$. Porém, um estimador de semivariograma não pode ser usado diretamente em predições ou interpolação espacial porque o mesmo não é necessariamente condicionalmente funções definidas positivas. A ausência dessa propriedade pode resultar em problemáticos erros quadráticos médios de predição negativos, se a curva (função) ajustada não atender a essa condição.

Assim, a estimação de θ pode ser considerada uma tarefa de ajustamento de curvas e, frequentemente, esse trabalho de ajuste é realizado a sentimento, sem nenhum critério formal de estimação. Entretanto, com essa finalidade existem os critérios de ajustes de quadrados mínimos que são baseados nos estimadores de semivariogramas. Portanto, os métodos mais utilizados são os métodos de quadrados mínimos ordinários (QMO) e os quadrados mínimos ponderados (QMP). Ainda, estão disponíveis os critérios da máxima verossimilhança (MV) e máxima verossimilhança restrita (MVR), os quais são métodos paramétricos de ajustes aplicados diretamente sobre a massa de dados. Além desses, têm-se também as abordagens bayesianas. Cada um desses métodos são particularmente abordados a seguir, exceto os métodos bayesianos (DIGGLE; RIBEIRO JUNIOR, 2007; CRESSIE,

1993).

2.4.1 Estimação por máxima verossimilhança

O procedimento de estimação baseado na máxima verossimilhança (MV) exige que a distribuição de um campo aleatório espacial seja conhecida e repouse terminantemente sobre a suposição de o processo ser um campo aleatório gaussiano (SCHABENBERGER; GOTWAY, 2005). Segundo Waller e Gotway (2004, p. 286), se os dados, $Z(x)$, seguem uma distribuição gaussiana multivariada, com média $\mu(x)$ ou, simplesmente, 1μ (em que 1 é um vetor de $1's$) e matriz de variância-covariância $\Sigma(\theta)$, as técnicas baseadas em verossimilhança podem ser utilizadas para estimar θ. Então, maximizando a verossimilhança multivariada em relação a θ produz o estimador de máxima verossimilhança (EMV) de θ. Mas, nesta pesquisa, será descrito o caso em que a média é constante, $E[Z(x)] = 1\mu$, combinada com a suposição de estacionaridade intrínseca ou de segunda ordem. Obviamente, a estimação por verossimilhança não impõe essa restrição sobre a média e, portanto, pode ser relaxada para atender àqueles casos em que existe uma deriva no contexto espacial.

O método de estimação baseado em verossimilhança possui um lugar de destaque na Estatística clássica e, atualmente, tem sido um poderoso aliado para ajustar modelos ou estimar os parâmetros de covariância nas análises da Geoestatística moderna (DIGGLE; RIBEIRO JUNIOR, 2007). Embora seja um método largamente difundido no meio estatístico de modo geral, pode-se dizer que o mesmo é um recurso ainda incipiente nas aplicações da Estatística Espacial como um todo. A teoria central da verossimilhança está fundamentada no conceito da função de verossimilhança e será definida a seguir.

Para dados correlacionados normalmente distribuídos a função de verossi-

milhança é

$$L(\beta,\theta \mid Z(x)) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\mid \Sigma \mid^{1/2}} e^{-\frac{1}{2}(Z(x)-V\beta)'\Sigma(\theta)^{-1}(Z(x)-V\beta)}, \qquad (2.9)$$

em que as variâncias e as covariâncias pode ser estimadas por $\widehat{Var}[Z(x)] = \Sigma(\theta)$ (CHABENBERBER; GOTWAY, 2005; GUMPERTZ, 1999).

Suponha que os dados $Z(x)$ são gaussianos e, ainda, admitindo-se uma especificação linear para a tendência espacial, $\mu(x_i)$, com $i = 1,\ldots,n$, o que permite introduzir uma superfície de tendência polinomial por meio da seguinte relação $\mu(x_i) = V\beta$. Consequentemente, tem-se que $Z(x) \sim GAU(V\beta,\Sigma(\theta))$, em que V é uma matriz de covariáveis $n \times p$ de posto $p < n$, β corresponde ao vetor de parâmetros de regressão, e a matriz $\Sigma(\theta) = \text{Cov}[Z(x_i),Z(x_j)]$, $n \times n$. Assim, a função log-verossimilhança é dada por

$$\begin{aligned} \ell(\beta,\theta \mid Z(x)) = &- \left(\frac{n}{2}\right) ln(2\pi) - \left(\frac{1}{2}\right) ln \mid \Sigma(\theta) \mid \\ &- \left(\frac{1}{2}\right) (Z(x) - V\beta)'\Sigma(\theta)^{-1}(Z(x) - V\beta), \end{aligned} \qquad (2.10)$$

com $\beta \in \mathbb{R}^p, \theta \in \Theta$ (denominado espaço paramétrico de θ) e os estimadores de MV dos parâmetros $\hat{\beta}$ e $\hat{\theta}$ devem satisfazer a

$$L(\hat{\beta},\hat{\theta}) = inf\{L(\beta,\theta) : \beta \in \mathbb{R}^p, \theta \in \Theta\}.$$

De acordo com Azzalini (1996 citado por ARBIA 2006, p. 81), os estimadores de máxima verossimilhança satisfazem um conjunto de propriedades ótimas sob as hipóteses da regularidade de modelos de probabilidade e a de independência de modelos amostrais. Este autor também cita que Bates e White (1985) e Heijmans e Magnus (1986) demostraram que as estimativas de máxima verossi-

milhança mantêm suas propriedades, mesmo que elas seja baseadas em obervações retiradas de processos aleatórios com componentes não independentes. Além disso, as seguintes condições são asseguradas:

(i) $L(\beta,\theta \mid Z(x))$ existe

(ii) as seguintes derivadas existem $\forall\theta \in \Theta$

$$L'(\theta) = \frac{\partial L(\theta, Z(x))}{\partial\theta} \qquad L''(\theta) = \frac{\partial^2 L(\theta, Z(x))}{\partial^2\theta} \qquad L'''(\theta) = \frac{\partial^3 L(\theta, Z(x))}{\partial^3\theta}$$

(iii) $L'(\theta) = \frac{\partial L(\theta,Z(x))}{\partial\theta} < \infty$

(iv) Σ existe e é não-singular, com $\Sigma = \{\gamma(x_i,x_j)\}$ a matriz de variância-covariância do campo aleatório gerando os dados.

2.4.2 Estimação por máxima verossimilhança restrita

Embora os estimadores de MV apresentem ótimas propriedades estatísticas, existem algumas desvantagens na aplicação do método, a saber: i) é necessário que a distribuição dos dados seja conhecida e ii) os estimadores de MV são viesados, pois os efeitos fixos são assumidos conhecidos. No contexto de estimação de componentes de covariância, em delineamentos experimentais, por exemplo, esse viés é decorrente da perda dos graus de liberdade no ajuste dos efeitos fixos, isto é, deve-se sacrificar uma observação ($Z(x)$ tem $(n-1)$ elementos) para estimar θ com esse método.

Com o objetivo de remover esse viés, Patterson e Thompson (1971) introduziram o método da máxima verossimilhança restrita (MVR), o qual considera apenas a parte da função de verossimilhança que independe dos efeitos fixos. Desse modo, esses mesmos autores sugeriram dividir cada observação $Z(x_i)$ em duas partes independentes, uma referente aos efeitos fixos e outra aos efei-

tos aleatórios. Basicamente, a ideia consiste em obter os estimadores de MVR aplicando-se a máxima verossimilhança aos contrastes de erros em vez de aplicá-lo na massa de dados propriamente dita. Esse critério (MVR) também foi denominado por Rao (1979) como método da máxima verossimilhança marginal (MML) e, mais recentemente, é chamado por alguns autores de máxima verossimilhança residual, mas mantiveram a sigla MVR. Segundo Cressie (1993), uma combinação linear $a'Z(x)$ é denominada um contraste de erro, com $E[a'Z(x)] = 0$, para todo $\beta \in \mathbb{R}^p$ e $\theta \in \Theta$. Realmente, $a'Z(x)$ é um constraste de erro se e somente se $a'V = 0'$.

Nesse caso, seja o processo espacial $Z(x) = Z$ e, assumindo-se o modelo $E[Z] = V\beta = 1'\mu$, é possível aplicar a transformação linear $\tilde{Z} = BZ$ nos dados de modo que a distribuição de $\tilde{Z}$ não dependa de β; isto é, $B'V\beta = 0$. Esta restrição imposta garante a não-dependência dos dados trasformados do parâmetro β, porém, causa efetiva redução da dimensão n de $\tilde{Z}$, modificando-a para a dimensão $n - p$, sendo p o número de elementos de β. Partindo dessa formulação, o critério MVR estima o valor dos parâmetros $\theta = (\tau^2, \sigma^2, \varphi)$, os quais determinam a estrutura de covariância dos dados, pela máxima verossimilhança aplicada aos dados transformados $\tilde{Z}$ (DIGGLE; RIBEIRO JUNIOR, 2007). De fato, sempre é possível buscar uma conveniente matriz B, mesmo desconhecendo os verdadeiros valores dos parâmetros θ ou β. Isso pode ser efetuado pela projeção residual de quadrados mínimos ordinários, ou seja,

$$B = I - V(V'V)^{-1}V'.$$

É importante observar que pelo fato de $\tilde{Z}$ ser uma transformação linear de Z, este preserva a propriedade de ser uma distribuição gaussiana assumida nos dados.

Agora, considere $\omega = BV$ um vetor de $n - p$ (p é posto de V) contrastes

de erros linearmente independentes; isso implica que as $n - p$ colunas de B são linearmente independentes e $B'V = 0$, sob a suposição gaussiana (2.10), $\omega \sim Gau(0, B'\Sigma(\theta)B)$, o qual não depende de β. Então, de acordo com Cressie (1993), a função generalizada log-verossimilhança negativa é

$$L_\omega(\theta) = \left(\frac{n-p}{2}\right) log(2\pi) + \frac{1}{2} log \mid B'\Sigma(\theta)B) \mid + \frac{1}{2}\omega'(B'\Sigma(\theta)B)^{-1}\omega. \tag{2.11}$$

Caso outro conjunto de $(n - p)$ contrastes linearmente independentes formem o vetor ω, obtém-se a nova função log-verossimilhança negativa que difere de L_ω apenas adicionando-se uma constante, por Harville (1974) citado em Cressie (1993). Realmente, para a matriz B que satisfaz $BB' = I - V(V'V)^{-1}V'$ e $BB' = I$,

$$L_{\boldsymbol{\omega}}(\theta) = \left(\frac{n-p}{2}\right) log(2\pi) - \left(\frac{1}{2}\right) log \mid V'V \mid + \left(\frac{1}{2}\right) log \mid \Sigma(\theta) \mid + \left(\frac{1}{2}\right) log \mid (V'\Sigma(\theta)^{-1}V) \mid + \left(\frac{1}{2}\right) Z(x)'G(\theta)Z(x), \tag{2.12}$$

sendo $G(\theta) = \Sigma(\theta)^{-1} - \Sigma(\theta)^{-1}V(V'\Sigma(\theta)^{-1}V)^{-1}V'\Sigma(\theta)^{-1}$. Então, uma estimativa MVR de θ é obtida minimizando (2.12) em relação à θ.

Certamente, a principal diferença entre os métodos MV e MVR é que o primeiro usa a função de verossimilhança de $z(x)$ ou o *log* desta função, enquanto o segundo adota a função de verossimilhança de um conjunto de contrastes de erros, ω, com $E[\omega] = 0$, o qual representa efetivamente as observações transfomadas. Cressie (1993) chama a atenção para outra importante distinção entre os estimadores MVR e MV, quando o valor de p é maior que o de n. Neste

caso, o critério MVR não gera subestimativas de θ como ocorre frequentemente com MV. Consequentemente, supondo um modelo com tendência (média espacialmente não-constante), as variâncias de predições dos preditores lineares da Geoestatística (krigagem universal) obtidas por meio dos estimadores MVR de θ são, geralmente, menos viesadas que aquelas obtidas por MV.

2.4.3 Método dos quadrados mínimos e o semivariograma

Os métodos de ajustes baseados em máxima verossimilhança vistos anteriormente ignoram completamente o gráfico do semivariograma estimado, $\hat{\gamma}(h)$, e ajustam uma curva de semivariograma teórico, $\gamma(h;\theta)$, análoga àquele estimado, a partir dos dados originais. Já os métodos dos quadrados mínimos modelam a dependência espacial diretamente do semivariograma experimental (CHABENBER; GOTWAY, 2005, p. 164). Essas técnicas são procedimentos de ajustes de curvas bastante conhecidas pelos estatísticos e já são consagradas nas aplicações de modelos lineares, em geral. Claro que, indiferente do método de ajuste a adotar, é altamente recomendável confrontar graficamente o semivariograma experimental (estimado) com a curva do semivariograma teórico ajustado. Esse procedimento, embora visual, é uma ferramenta exploratória de grande valia. Cressie (1993) afirma que esse procedimento é uma ferramenta de diagnóstico inestimável. Na prática Geoestatística, os parâmetros de covariância podem ser estimados ajustando-se uma função de covariância (somente funções autorizadas) a um semivariograma experimental, o qual serve de base de dados (ou pseudo-dados) para esse processo (SCHABENBERGER; GOTWAY, 2005). Por exemplo, considere um estimador de semivariograma definido por k incrementos de distância h (ou k *lags*), então, um modelo de semivariograma $\gamma(h,\theta)$ pode ser ajustado aos pseudo-

dados

$$\hat{\gamma}(h) = \hat{\gamma}(h_1), \ldots, \hat{\gamma}(h_k).$$

Como, geralmente, o relacionamento entre $\hat{\gamma}(h)$ e h é não linear, uma regressão de quadrados mínimos não linear pode ser usada para estimar θ. Assim, o método de quadrados mínimos ordinários (QMO), não linear, encontra o valor de $\hat{\theta}$ minimizando a distância vertical quadrática entre os semivariogramas empírico e teórico, isto é, minimizando (WALLER; GOTWAY, 2004, p. 285)

$$S_o(\theta) = \sum_{k=1}^{m} [\hat{\gamma}(h_k) - \gamma(h_k;\theta)]^2, \tag{2.13}$$

para uma dada direção, considerando um número m de distância máxima.

No entanto, as estimativas $\hat{\gamma}(h)$ são correlacionadas e têm diferentes variâncias (variância heterocedástica), violando as suposições gerais dos QMO e, por isso, esse método não formece o melhor estimador de θ (CRESSIE, 1985; GUMPERTZ, 1999; ARBIA, 2006). Assim, a ação apropriada quando as observações são correlacionadas e heterocedástica é o usual critério dos quadrados mínimos generalizados (QMG) e a função objetivo a ser minimizada é (CRESSIE, 1985):

$$S_g(\theta) = (\hat{\gamma}(h) - \gamma(h,\theta))'\mathbf{R}(\theta)^{-1}(\hat{\gamma}(h) - \gamma(h,\theta)). \tag{2.14}$$

O método dos quadrados mínimos não fazem suposições distribucionais sobre $\hat{\gamma}(h)$ a partir dos dois primeiros momentos. Esses métodos são formalizados pelo modelo

$$\gamma(h) = \gamma(h,\theta) + \varepsilon(h), \tag{2.15}$$

em que $\gamma(h,\theta) = [\gamma(h_1,\theta), \ldots, \gamma(h_k,\theta)]'$. Assume-se que o vetor de erros, $\varepsilon_{k\times 1}$, nesse modelo tem média zero (isto é $E[\varepsilon(h)] = 0$) e a matriz variância-covariância

dos erros dada $Var[\varepsilon(h)] = \mathbf{R}(\theta)$. Note que $\mathbf{R}(\theta)$ depende do desconhecido parâmetro θ e, portanto, esse estimador é calculado iterativamente (por meio do algoritmo de Gauss-Newton), a partir de valores iniciais que são melhorados (ou atualizados) até a função objetivo ser minimizada (SCHABENBERGER; GOTWAY, 2005).

Sob a suposição de que $Z(x)$ é um campo aleatório gaussiano, Cressie (1985) mostrou que a variância do estimador de Matheron (isto é, a matriz de covariância), $\hat{\gamma}(h)$, é dada pela expressão,

$$Var[\gamma(h)] = \frac{2[\hat{\gamma}(h)]^2}{N(h)}\left(1 + \frac{1}{N(h)}\sum_{i=1}^{N(h)}\sum_{i\neq j}\left(\frac{\gamma(z_i - z_j - h) + \gamma(z_i - z_j + h) - 2\gamma(z_i - z_j)}{2\gamma(h)}\right)^2\right). \tag{2.16}$$

Devido à dificuldade de se minimizar a soma de quadrados generalizada e, então, obter os elementos de $\mathbf{R}(\theta)$, Cressie (1985) propõe uma aproximação para as entradas da diagonal de $\mathbf{R}(\theta)$ tal como

$$Var[\gamma(h_k)] \approx 2\frac{\gamma(h,\theta)^2}{N(h_k)}. \tag{2.17}$$

Entretanto, essa fórmula da variância é apropriada se a quantidade $[Z(x_i) - Z(x_j)]^2$ for independente e a suposição gaussiana é assugurada.

Outra possibilidade de se ajustar o semivariograma é o método de ajuste por quadrados mínimos ponderados (QMP). Essa abordagem considera somente as entradas da diagonal de $\mathbf{R}(\theta)$ como forma de pesos para efetuar o ajuste do modelo. Desse modo, o método dos QMP substitui $\mathbf{R}(\theta)$ — quadrados mínimos generalizados — pela matriz diagonal $\mathbf{W}(\theta)$ cujas entradas são dadas em (2.17) (SCHABENBERGER; GOTWAY, 2005). Cressie (1985) mostrou que o estimador

de QMP baseado em (2.17) tem um bom desempenho na prática.

O critério QMP de ajuste de um modelo de semivariograma pode ser implementado por qualquer algoritmo de estimação não-linear. Sob um modelo gaussiano, em cada semivariograma experimental, tem-se que a $E[\gamma(h)] = \gamma(h;\theta)$ e a variância é dada por $2\gamma(h;\theta)^2$. Com essa formulação, Cressie (1993) propôs o critério QMP para estimar o parâmetro θ de um modelo de semivariograma por minimizar a função

$$S_w(\theta) = (\hat{\gamma}(h) - \gamma(h,\theta))'\mathbf{W}(\theta)^{-1}(\hat{\gamma}(h) - \gamma(h,\theta)) \tag{2.18}$$

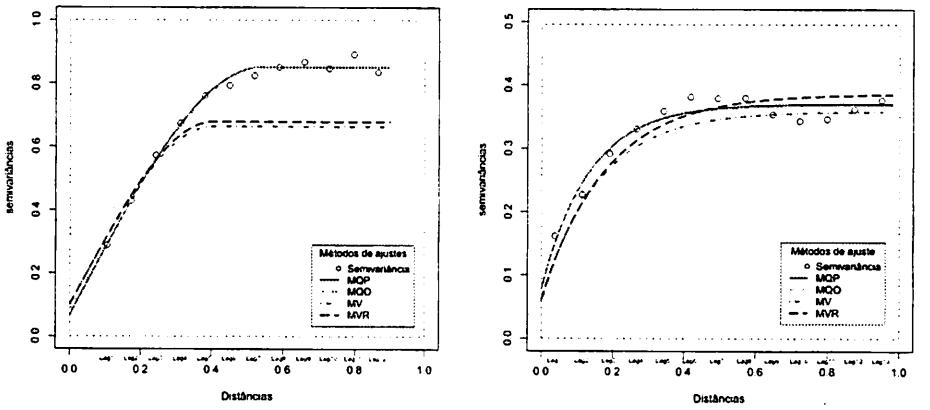
$$= \sum_{k=1}^{m} \frac{N(h_k)}{[2\gamma(h_k;\theta)]^2}[\hat{\gamma}(h_k) - \gamma(h_k;\theta)]^2. \tag{2.19}$$

Segundo Cressie (1993, p. 99), o método QMP, em (2.18), é considerado um critério para ajustes de regressão não-linear ponderada com pesos $\frac{N(h_k)}{[\gamma(h_k;\theta)]^2}$, sendo $k = 1,\ldots,m$, com o qual se estima θ a partir de qualquer rotina de estimação não-linear iterativa, tal como o algoritmo de Gauss-Newton, por exemplo.

Na Figura 2.11 tem-se um exemplo de modelagem de um semivariograma experimental ilustrando um ajuste com o modelo esférico e outro com o exponencial, em que pode ser visualizado o comportamento (ou a qualidade do ajuste) de cada um dos critérios QMP, QMO, MV e MVR, considerando dados simulados.

2.5 Modelos teóricos de semivariogramas

Ajustar um modelo teórico ao semivariograma experimental ou empírico não é semelhante aos ajustes usuais em modelos de regressão. Em Geoestatística, o modelo que define as propriedades do padrão de dependência espacial, o qual é uma função de covariância usada em interpolações lineares por meio da krigagem, deve obedecer à condição de funções definidas positivas Searle (1982). Portanto,



(a) Modelo Esférico (b) Modelo Exponencial

Figura 2.11 Exemplo de ajuste de semivariograma experimental usando os critérios QMP, QMO, MV e MVR

a prática de modelagem geoestatística fica restringida apenas a um conjunto de modelos devidamente autorizados (apenas as funções definidas positivas). Não se pode considerar o modelo pelo simples fato de o mesmo se ajustar razoavelmente aos pontos do gráfico do semivariograma experimental. Por motivos óbvios, aqui serão apresentados apenas os modelos básicos: são modelos simples, isotrópicos (independente da direção). Também são os mais usados em modelagem de índice espacial contínuo. Convenientemente, os semivariogramas são classificados em dois tipos: aqueles que possuem patamar e os que não possuem patamar.

2.5.1 Semivariograma com patamar ou modelos de transição

São frequentemente referidos como modelos de transição. Esses tipos de modelos são bem comportados e possuem seus respectivos patamares bem definidos, caracterizando a estabilização de variância quando se atinge o alcance. Porém,

alguns desses modelos de transição alcançam seu patamar assintoticamente. Para esses modelos, em especial, o alcance é determinado, arbitrariamente, com sendo a distância na qual se alcança 95% do patamar. De maneira bem prática, o alcance é definido como sendo a distância que separa as variáveis aleatórias correlacionadas daquelas não correlacionadas. Embora, na prática, os cálculos que definem o padrão de correlação espacial sejam realizados por meio da matriz de covariâncias espaciais, esses modelos de funções de covariâncias são exibidos em gráficos, preferivelmente, por meio dos semivariogramas, e não pelo modelo da função de covariância (covariograma ou correlograma) propriamente dito. A relação que permite essa mudança teórica da estrutura de covariância para o semivariograma é dada pela expressão $\gamma(h) = C(0) - C(h)$. A seguir serão abordados os principais modelos teóricos de semivariogramas que possuem patamar definido.

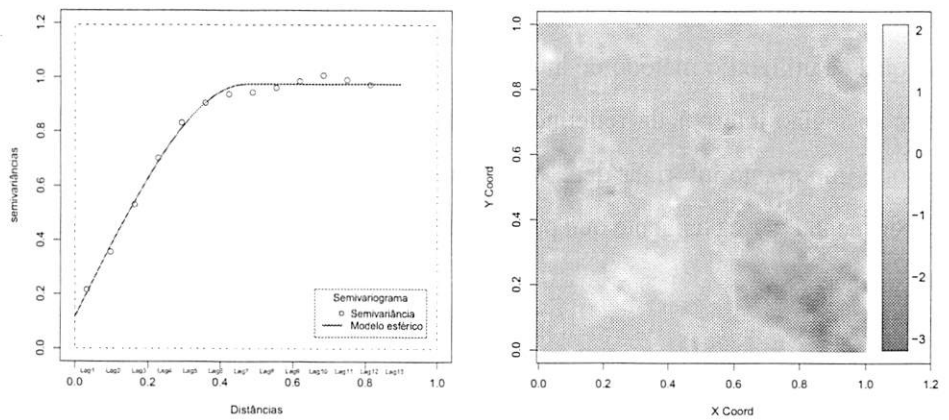
Modelo esférico

Particularmente, trata-se de um dos modelos mais usados em Geoestatística. É uma função descrita apenas pelos parâmetros alcance e patamar. Esse modelo destaca-se dos demais porque o patamar alcançado pelo modelo é definitivamente real; em outras palavras, o patamar não tem comportamento assintótico. Na Figura 2.12 observa-se um exemplo do ajuste do semivariograma empírico por meio de um modelo teórico esférico e o mapa gerado, a partir desse ajuste, caracterizando o padrão espacial. A equação para cálculo das semivariâncias é dada

pela equação,

$$\gamma(h;\tau^2,\varrho^2,\varphi)_{esf} = \begin{cases} 0 & h = 0 \\ \tau^2 + \varrho^2\left(\frac{3}{2}\cdot\frac{h}{\varphi} - \frac{1}{2}\cdot(\frac{h}{\varphi})^3\right) & 0 < h \leq \varphi, \\ \tau^2 + \varrho^2 & \varphi > 0 \end{cases} \tag{2.20}$$

em que o parâmetro τ^2 define o efeito pepita, φ indica o alcance, ϱ^2 é a contribuição e vetor h representa a distância. O modelo esférico tem um comportamento linear para pequenas distâncias mais próximo da origem do gráfico, e para maiores distâncias tem uma tendência curvilínea, quando o *lag* se aproxima do alcance atingindo o patamar. Deve-se chamar a atenção do leitor que, ao ajustar este modelo ao semivariograma experimental, a reta tangente à origem intercepta o patamar a dois terços do alcance ($\sigma^2 = \frac{2}{3}\varphi$).



(a) Ajuste do semivariograma

(b) Padrão espacial do modelo esfério

Figura 2.12 Modelo esférico ajustado pelo método de quadrados mínimos ponderados e o mapa do padrão espacial refletido por esse modelo

Modelo exponencial

Esse também é um modelo de transição bastante usado na prática e a equação usada para modelagem dos processos estocásticos é dada por:

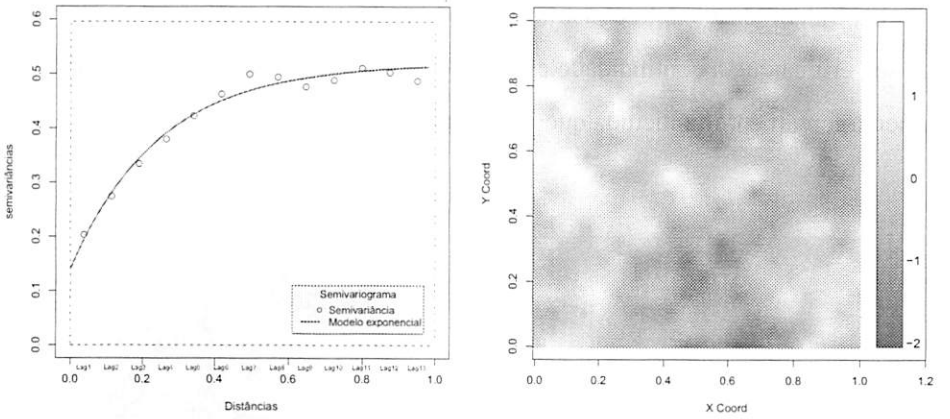
$$\gamma(h;\tau^2,\varrho^2,\varphi)_{exp} = \begin{cases} 0 & h = 0 \\ \tau^2 + \varrho^2\left(1 - e^{-\frac{3h}{\varphi}}\right) & h \neq 0. \end{cases} \tag{2.21}$$

Note que os parâmetros $\{\tau^2, \varrho^2, \varphi\}$ possuem a mesma nomenclatura definida para o modelo esférico na Equação (2.20).

O modelo em questão atinge o patamar assintoticamente. Assim, o alcance prático do semivariograma é obtido tomando-se 95% do patamar. Semelhante ao modelo esférico, o modelo exponencial é linear para pequenas distâncias. No entanto, à medida que a distância aumenta, seu traçado se torna mais acentuado que o modelo esférico e, depois, seu gráfico segue suavemente buscando o platô, assintoticamente. Na Figura 2.13 observa-se comportamento do modelo exponencial no qual se utilizou o método de ajuste dos QMP, além de mostrar também a imagem da região interpolada refletindo o seu padrão espacial característico. Nesse caso, é importante informar que, ao se ajustar este modelo a um semivariograma experimental, a reta tangente que parte da origem intercepta o patamar a um quinto do alcance.

Modelo gaussiano

É o modelo de transição que tem a característica de modelar fenômenos extremamente contínuos. Sua equação é definida como:



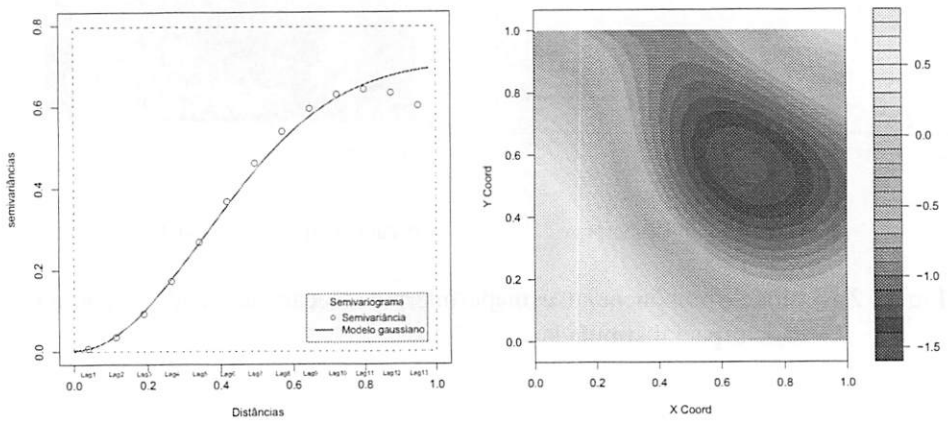
(a) Ajuste do semivariograma (b) Padrão espacial do modelo exponencial

Figura 2.13 Modelo exponencial e mapa interpolado caracterizando a continuidade espacial simulada

$$\gamma(h;\tau^2,\varrho^2,\varphi)_{gau} = \begin{cases} 0 & h = 0 \\ \tau^2 + \varrho^2\left(1 - e^{-3\left(\frac{h}{\varphi}\right)^2}\right) & h \neq 0, \end{cases} \tag{2.22}$$

sendo que os parâmetros $(\tau^2,\varrho^2,\varphi)$ seguem a terminologia dos modelos anteriores. Muito semelhante ao modelo exponencial, o modelo gaussiano também atinge o patamar assintoticamente, isto é, o que se obtém na verdade é o alcance prático cujo valor é 95% do patamar. Esse modelo é diferenciado pela sua forma parabólica próxima da origem, a qual potencializa o modelo em caracterizar suaves variações de padrões espaciais. Entre os modelos de transição o modelo gaussiano é o único que possui ponto de inflexão. Na Figura 2.14b observa-se o mapa da continuidade espacial gerado pelo modelo gaussiano ajustado, a partir de dados simulados, também ajustado pelo método dos QMP, conforme mostrado na Figura 2.14a. É importante destacar, nesse momento, a habilidade de o modelo gaussiano

representar ou modelar fenômenos altamente contínuos, conforme já mencionado. Essa habilidade do modelo é justificada pela sua forma parabólica na origem, em que se verifica que a continuidade espacial diminui lentamente (ou a variabilidade aumenta lentamente) à medida que h aumenta (Figura 2.14a).



(a) Ajuste do semivariograma (b) Padrão espacial do modelo Gaussiano

Figura 2.14 Mapa de contorno exemplificando o padrão espacial do modelo gaussiano ajustado pelo método de quadrados mínimos ponderado: padrão de dependência espacial altamente contínuo

Modelo efeito pepita puro

Quando, por meio de simples inspeção da variografia experimental, for observado um conjunto de pontos oscilando aleatoriamente em torno de um valor constante (o patamar, por exemplo), diz-se que o fenômeno apresenta um modelo efeito pepita puro (Figura 2.15a). Colocando em outras palavras, a julgar apenas pela análise descritiva visual, esse tipo de comportamento denuncia que as amostras são originadas de uma população, em geral, desconhecida, que, provavelmente

não possui a qualidade de estar estruturada no espaço — qualquer semelhança entre amostras vizinhas é mero fruto do acaso. Portanto, não existe dependência espacial para ser modelada e, consequentemente, nesse sentido, é impossível se fazer predições pontuais ou globais ou, ainda, construir mapas utilizando os usuais métodos de krigagem. Na Figura 2.15b mostra-se uma situação que ilustra bem esse comportamento, em que claramente o modelo gerou uma superfície repleta de pontos aleatório, sem padrão espacial definido. Esse modelo também pertence à classe dos modelos de transição, como o esférico, por exemplo. Nesse caso, a própria média pode ser usada para representar o fenômeno e as técnicas usuais da estatística convencional podem ser aplicadas normalmente, não se considerando a componente espacial, é claro. Contudo, mesmo na ausência de dependência espacial, mapas ainda podem ser construídos por meio de outros métodos de interpolação, tais como os métodos da triangulação, dos polígonos, inverso da distância etc. (WEBSTER; OVLIVER, 2001; ISAAKS; SRIVASTAVA, 1989).

O modelo efeito pepita puro pode ser explícito conforme a equação:

$$\gamma(h;\tau^2)_{pep} = \begin{cases} 0 & \text{para } h = 0, \\ \tau^2 & \forall \quad h > \varphi. \end{cases} \tag{2.23}$$

De fato, um semivariograma com essa característica informa que há indício de total ausência de dependência espacial entre quaisquer pares de variáveis aleatórias $Z(x_i)$ e $Z(x_i + h)$. O comportamento do modelo em pauta $\gamma(h;\tau^2)_{pep}$ é praticamente devido à ocorrência de um alcance (φ) muito pequeno quando comparado com maiores distâncias h entre os pontos amostrais da variável regionalizada, isto é, a escala na qual está ocorrendo a dependência espacial é menor que a escala de quaisquer pares de distância (relativo ao espaçamento) entre os pontos amostrados $Z(x_i)$ e $Z(x_i + h)$ elaborados ou determinados no planejamento

amostral e, portanto, não será detectada. Segundo Journel e Huijbregts (1991, p. 152), o efeito pepita é usado para caracterizar a influência residual de todas as variabilidades nas quais têm alcances muito menores que as distâncias disponíveis de observações. É óbvio e oportuno informar que o efeito pepita está diretamente relacionado com a qualidade do planejamento amostral, exceto os casos em que realmente não há dependência espacial entre os pares de amostras. Cada estudo de caso exige perícia do pesquisador para que sejam elaborados planos amostrais eficientes e capazes de detectar otimamente a correlação espacial. Caso contrário, o variograma não será capaz de caracterizar com precisão o tipo de dependência espacial, se ela existir ou, ainda, ter-se-á um efeito pepita expressivo, o que não é desejável porque, desse modo, o modelo ajustado expressará um menor grau de dependência espacial que o devido (CAMBARDELLA et al., 1994). Além disso, um plano amostral inadequado não permite que o semivariograma descreva a continuidade espacial que possa estar ocorrendo em micro, pequena ou média escala.

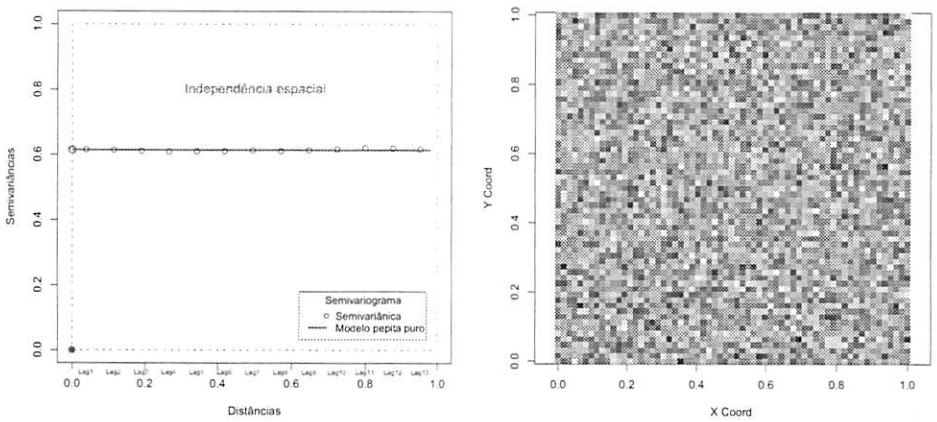
Percebe-se, pela Figura 2.15, que foi o efeito pepita que gerou uma descontinuidade na origem do semivariograma, em que para $h = 0$, o gráfico teve um intercepto diferente de zero (0). Em síntese, o efeito pepita pode ser justificado como sendo erro de amostragem e, portanto, pode ser minimizado quando se elabora um "bom" plano de amostragem.

2.5.2 Modelos de semivariogramas sem patamar

Essa classe de modelos corresponde a fenômenos que têm uma capacidade infinita para dispersão espacial e, por isso, a variância não é finita e, consequentemente, a função de covariância ou de correlação não pode ser definida. Tais modelos estão ilustrados na Figura 2.16 e serão definidos a seguir.

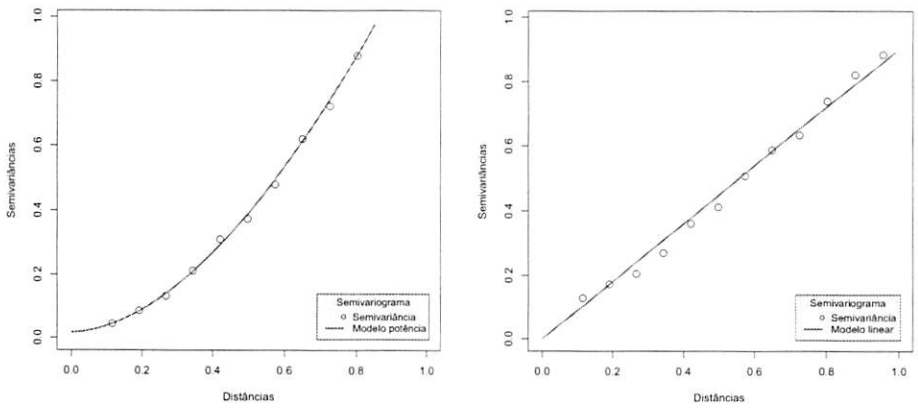
Modelo potência

Trata-se de um tipo de modelo que possui dispersão infinita e, portanto, o semivariograma não atinge o patamar. Por esse motivo, os parâmetros alcance



(a) Ajuste do semivariograma (b) Imagem do efeito pepita puro

Figura 2.15 Modelo efeito pepita puro e seu mapa característico



(a) Modelo de potência (b) Modelo linear

Figura 2.16 Exemplos de modelos teóricos sem patamar

(φ) e patamar (σ^2) não podem ser interpretados como nos modelos teóricos de transição e, nesse caso, deve-se assumir que o processo é somente intrinsecamente estacionário. A função do semivariograma é dada por:

$$\gamma(h;\tau^2,w,\phi)_{pot} = \begin{cases} 0 & \text{para} \quad h = 0, \\ \tau^2 + w \cdot h^{\phi} & \forall \quad h \neq 0, \end{cases} \tag{2.24}$$

com $0 \leq \phi < 2$, $\tau^2 \geq$ e $w \geq 0$, sendo que ϕ descreve a curvatura e os parametros τ^2 e w descrevem o efeito pepita e a intensidade da variação, respectivamente. Se $\phi \geq 2$, o modelo viola a hipótese intrínseca (SCHABENBERGER; GOTWAY, 2005).

Visto que o correspondente processo não possui estacionaridade de segunda ordem, o semivariograma não pode ser construído pela função de covariância. Nota-se na Figura 2.16a, que o modelo potência claramente não satisfaz a hipótese de estacionaridade de segunda.

Modelo linear

O modelo linear também não é um modelo de transição e trata-se de um caso particular do modelo potência com $\varphi = 1$. Percebe-se que o modelo linear também não alcança o patamar, mas cresce linearmente com o vetor distância h. Sua representação gráfica encontra-se na Figura 2.16b, em que se pode observar a ausência de um patamar, o que caracteriza que o modelo possui variância infinita. A forma padronizada do modelo pode ser definida simplesmente como:

$$\gamma(h;\tau^2,w)_{lin} = \begin{cases} 0 & \text{para} \quad h = 0, \\ \tau^2 + w \cdot h & \forall \quad h \neq 0, \end{cases} \tag{2.25}$$

com $\tau^2 \geq 0$ e $w \geq 0$.

O modelo de semivariograma linear não é estritamente positivo definido porque existem combinações de pesos no processo de cálculo do semivariograma que permite ocorrer valores negativos. Porém, estabelecendo-se a condição de que os pesos obtidos no procedimento de estimação devam somar zero, garantirá que o modelo seja válido para a predição por krigagem ordinária.

2.6 Outros modelos gerais de semivariogramas

Nesta seção serão apresentados, de forma sucinta, alguns modelos teóricos de semivariogramas que estão disponíveis aos especialistas em Geoestatística. Atualmente, existem em torno de 33 modelos teóricos de semivariogramas disponíveis e muitos deles estão implementados em diversos sofwares (livres ou comerciais). Entretanto, dar-se-á ênfase ao programa estatístico gratuito R (R DEVELOPMENT CORE TEAM, 2008). É de suma importância destacar os pacotes computacionais RandomFields (SCHLATHER, 2010), geoR (RIBEIRO JUNIOR; DIGGLE, 2001), spatial (VENABLES; RIPLEY, 2002), geoRglm (CHRISTENSEN; RIBEIRO JUNIOR, 2011) e gstat (PEBESMA, 2004), que são poderosas ferramentas para análise geoestatística (estatística espacial) no R. Esses pacotes reunidos contemplam os mais avançados recursos para modelagem e predição geoestatística e, portanto, o R está à frente de qualquer outro software especializado em analisar dados geoestatísticos. Indiscutivelmente, o R fornece a maior variedade de modelos teóricos de semivariogramas para se realizar ensaios em ajustes de semivariograma experimental, bem como realizar predições. A seguir serão descritos apenas alguns desses modelos de semivariograma, que estão ilustrados na Figura 2.17.

Modelo wave

É um modelo utilizado para representar fenômenos contínuos com periodicidade e sua expressão matemática é dada por:

$$\gamma_(h;\varphi)wav = \left(\frac{h}{\varphi}\right) \cdot sen\left(\frac{h}{\varphi}\right), \tag{2.26}$$

em que a terminologia φ descreve o alcance e h representa distância *lag*.

O patamar do modelo wave é alcançado assintoticamente. O parâmetro alcance teórico é igual φ, sendo o alcance experimental igual a 3φ e tem comportamento parabólico na origem (Figura 2.17a).

Modelo k-Bessel (Matérn)

Um exemplo do modelo Matérn. Esse modelo é formalizado pela seguinte fórmula matemática:

$$\gamma(h;\tau^2,\varrho_k^2,\varphi_k,a)_{mat} = \begin{cases} 0, & h = 0, \\ \tau^2 + \varrho_k^2\left[1 - \frac{1}{2^{a-1}\Gamma(a)}\left(\frac{h}{\varphi_k}\right)^a K_a\frac{h}{\varphi_k}\right], & h > 0, \end{cases} \tag{2.27}$$

em que $\tau^2 \geq 0$, $\varrho_k^2 \geq 0$, $\varphi_k > 0$, $a \geq 0$, $K_a(\cdot)$ denota a função Bessel modificada do segundo tipo de ordem a (esta função estabelece a suavidade do processo) e $\Gamma(\cdot)$ é a função gama. Segundo Waller e Gotway (2004), essa família de modelos é denomidada pelos geoestatísticos de modelo k-Bessel, devido à sua dependência de $K_a(\cdot)$ e, recentemente, eles redescobriram a sua utilidade e o renomearam de modelos da classe Matérn, baseados na formalização inicial de Matérn (1960). Os modelos dessa família atinge o patamar assintoticamente. O comportamento típico desse modelo é mostrado na Figura 2.17b.

Modelo Cauchy

O modelo é formulado matematicamente pela expressão:

$$\gamma_(h;\varphi)cau = \left[1 + \left(\frac{h}{\varphi}\right)^2\right]^{-k}, \tag{2.28}$$

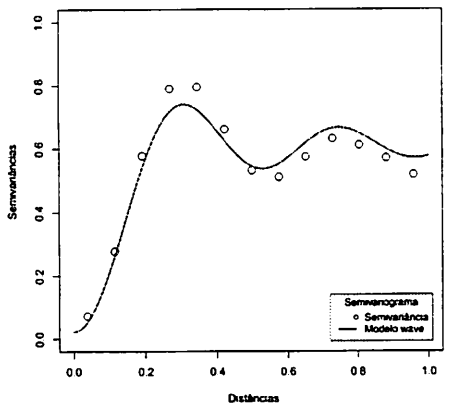
em que φ é o parâmetro da função de correlação, denominado patamar e k é o parâmetro suavisador do modelo em questão. Um exemplo gráfico do modelo Cauchy é visto na Figura 2.17c, em que se observa um comportamento parabólico do mesmo na origem, o que indica que esse modelo tem habilidade de detectar fenômenos bastantes contínuos a curtas distâncias e, depois, cresce linearmente buscando um patamar.

Modelo cúbico

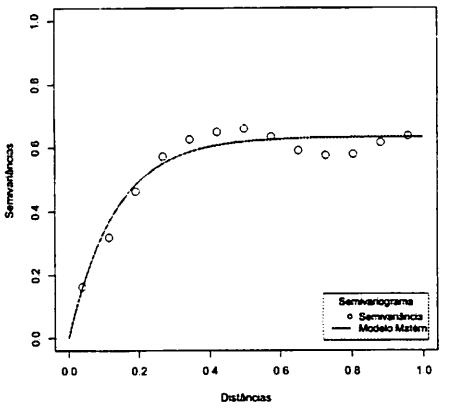
A expressão matemática do modelo pode ser formalizada com segue:

$$\gamma_(h;\varphi)cub = \begin{cases} 1 - \left[7\left(\frac{h}{\varphi}\right)^2\right] - 8{,}75\left[\left(\frac{h}{\varphi}\right)^3\right] + \\ +3{,}5\left[\left(\frac{h}{\varphi}\right)^5\right] - 0{,}75\left[\left(\frac{h}{\varphi}\right)^7\right] & h \leq \varphi \\ 0 & 0 < h < \varphi. \end{cases} \tag{2.29}$$

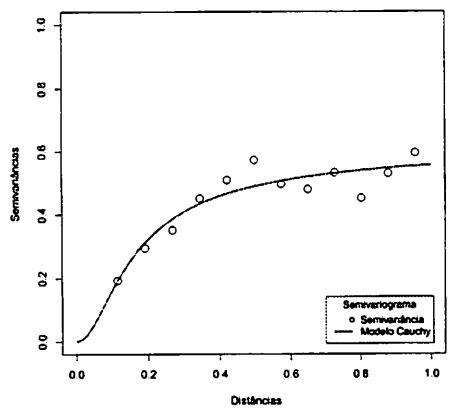
O comportamento desse modelo é bem parecido com o do modelo gaussiano, o que significa que este tem habilidade para representar fenômenos altamente contínuos. Esse fato pode ser constatado por sua forma parabólica na origem (Figura 2.17d).



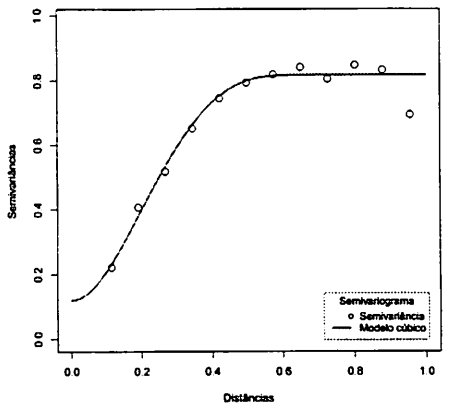
(a) Modelo teórico wave



(b) Modelo teórico Matérn



(c) Modelo teórico Cauchy



(d) Modelo teórico cúbico

Figura 2.17 Outros modelos teóricos de semivariograma

2.7 Propriedades intrínsecas do padrão espacial: isotropia, anisotropia

Evidentemente, há necessidade de se abordar o assunto sobre modelos com padrões espaciais isotrópicos, mas o tema anisotropia será tratado de modo superficial, somente para se tomar conhecimento da sua existência, bem como saber

identificar os tipos de anisotropia que, geralmente, ocorrem na natureza.

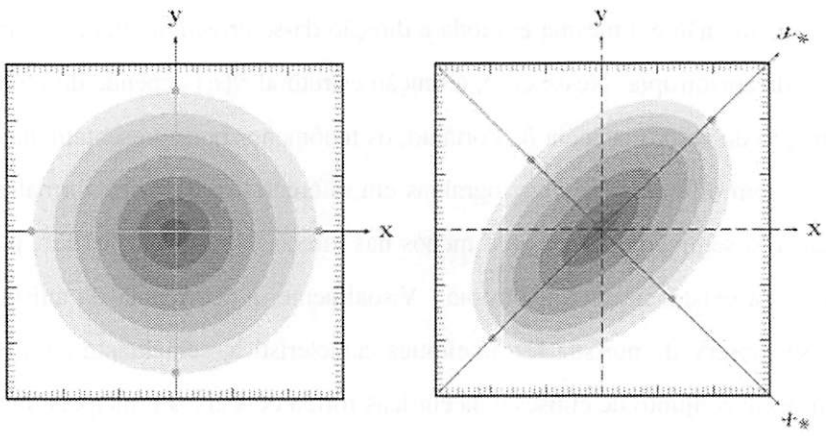
2.7.1 Isotropia e anisotropia

Quando o fenômeno possui o mesmo padrão de dependência espacial em qualquer direção dentro da região de estudo, diz-se que a característica estudada tem comportamento isotrópico. A isotropia pode ser facilmente verificada quando o semivariograma experimental é idêntico em todas as direções, isto é, quando a variabilidade da variável regionalizada, caracterizada pelo semivariograma $\hat{\gamma}(h)$. A Figura 2.18a expressa fielmente a característica de um comportamento isotrópico, pois as várias camadas circulares são um indicativo de que o fator direção, realmente, não provoca mudança nos parâmetros alcance e patamar.

Nos estudos de caracterização espacial em que a variabilidade da variável regionalizada não é a mesma em toda a direção dá-se origem a uma estrutura denominada anisotropia. Nesse caso, a função estrutural $\hat{\gamma}(h)$ depende do módulo e da direção do vetor distância h. Portanto, os fenômenos que apresentam anisotropia possuem diferentes semivariogramas em diferentes direções e, normalmente, estuda-se o semivariograma, pelo menos nas direções $0°, 45°, 90°$e$135°$, para se constatar a existência da anisotropia. Visualmente, a ocorrência da anisotropia pode ser observada por sua forma elíptica característica, semelhante à da Figura 2.18b. Esse conjunto de elipses concêntricas forma dois eixos principais de variabilidade espacial, com os quais se podem visualizar com certa precisão a intensidade de ocorrência, o impacto que a direção está provocando sobre a continuidade espacial e, também, a proporção entre a maior e a menor variabilidade à medida que se muda a direção. Assim, fica evidente que, nos fenômenos com efeito anisotrópico, a direção é fator determinante nas análises de continuidade espacial.

Na Figura 2.18b apresentam-se dois sistemas de coordenadas. Os eixos

(ox,oy) são condizentes com o sistema de coordenadas usado para georreferenciar os pontos amostrados. Isso implica em dizer que o eixo original é prefixado e não considera as possíveis variações espaciais que possam ocorrer nas diversas direções (anisotropia). Já os eixos (ox^*,oy^*) são obtidos pela rotação do eixo original no sentido de maior continuidade espacial, o qual considera as possíveis variações espaciais segundo a direção do vetor distância h. Esse procedimento é uma técnica usual que altera o sistema de coordenadas originais com o objetivo de transformar todo modelo de semivariograma anisotrópico em modelo isotrópico para, então, desenvolver toda análise estrutural e, depois, chegar ao objetivo final, que é o de realizar inferências sobre o fenômeno. Conclusivamente, toda a análise de continuidade espacial dos modelos de Geoestatística é baseada em modelos isotrópicos, portanto, há necessidade de se transformar as coordenadas originais.



(a) Isotropia (b) Anisotropia

Figura 2.18 Representação do comportamento espacial de variáveis contínuas. (a) Isotropia, representada pelos círculos concêntricos indicando mesmo padrão espacial em todas as direções; (b) Anisotropia, representada pelas elipses concêntricas, indicando que o padrão espacial depende da direção – o sistema de coordenadas original é rotacionado para o sistema (ox^*,oy^*) coincidindo com os eixos principais das elipses

A modelagem de estruturas espaciais anisotrópicas se processa por meio de um conjunto de técnicas que reduzem as diferentes estruturas de continuidade espacial existentes, nas diferentes direções, em um equivalente modelo isotrópico que seja capaz de representar o fenômeno. Há, basicamente, dois modelos mais comuns de anisotropia que devem ser referenciados: a anisotropia geométrica e a anisotropia zonal. Para mais detalhes sobre anisotropia, o consultar Webster e Oliver (2001); Chilès e Delfiner (1999); Isaaks e Srivastava (1989).

2.8 Metodos bootstrap

A teoria bootstrap é bastante extensa, portanto, não é de interesse dissertá-la totalmente nesta tese. Abordar-se os conceitos fundamentais do bootstrap, com algum relato histórico. Embora o bootstrap possibilite a construção de vários tipos de limites de confiança, apresenta-se somente o intervalo de confiança percentil. Caso seja necessário, pode-se consultar os seguintes autores: Chernick (2008); Davison e Hinkley (1997); Zoubir e Boashash (1998); Efron e Tibshirani (1993).

2.8.1 Visão geral

O paradigma bootstrap foi introduzido por Bradley Efron e sua teoria foi totalmente descrita no trabalho de Efron e Tibshirani (1993). A ideia central é a de gerar uma "grande quantidade" de amostras bootstrap, obtidas por amostragem aleatória simples, com reposição, a partir dos dados orignais. A metodologia bootstrap é uma abordagem não paramétrica de intensiva computação para inferência estatística que habilita estimar erros padrões por meio de uma adequada reamostragem de dados. Pode-se dizer que a ideia por trás do bootstrap é bem simples e já existe há pelo menos, dois séculos: é um método baseado no princípio

da amostragem com reposição.

Efron e Tibshirani (1993) abrem esse assunto, em seu livro, definindo a Estatística como uma ciência que se aprende a partir da experiência (ensaio) e que a mais antiga ciência de informação foi a Estatística, originariamente cerca de 1650. A partir daí, as técnicas estatísticas vêm se tornando métodos analíticos para tomada de decisões em diversas áreas do conhecimento.

De modo geral, a teoria estatística lida com três questões básicas:

1. Como eu poderia coletar meus dados?
2. Como eu poderia analisar e sumariar os dados que eu coletei?
3. Quão precisos são os resumos dos meus dados?

Essas três indagações estão relacionadas diretamente a uma parte da Estatística denominada inferência. O bootstrap é uma técnica relativamente recente, desenvolvida para se fazer um determinado tipo de inferência estatística. O seu desenvolvimento só se deu a partir da década de 1990 porque essa técnica requer o poder da informática moderna para simplificar ou contornar problemas teóricos intricados da estatística tradicional: o bootstrap é um método baseado em computador para avaliar medidas de acurácia para as estimativas estatísticas (EFRON; TIBSHIRANI, 1993).

2.8.2 Fundamentos básicos

Suponha que o conjunto de dados consiste de uma amostra aleatória de tamanho n a partir de uma distribuição de probabilidades desconhecida F sobre o conjunto dos números Reais,

$$Z_1, Z_2, \ldots, Z_n \sim F.$$

Seja, ainda, o vetor $z = (Z_1 = z_1, Z_2 = z_2, \ldots, Z_n = z_n)$ uma realização de n números observados de forma aleatória. Note que o termo "aleatório" significa dizer que os $Z_i's$ são variáveis aleatórias independente e identicamente distribuídas ($i.i.d.$), geradas pela função aleatória subjacente F_n. Denota-se também θ um parâmetro (ou característica) desconhecido inerente à função F, tais como média, variância, coeficiente de correlação, quantis ou qualquer estatística de particular interesse, por exemplo. Assim, a principal motivação aqui é encontrar o estimador de θ, bem como a distribuição de $\hat{\theta}$, com base na realização z. Essa descrição tem grande relevância prática quando há a necessidade de inferir sobre θ mediante apenas o conhecimento da característica amostral $\hat{\theta}$. Então, a arte de se encontrar a distribuição de $\hat{\theta}$ ou sua característica consiste em repetir o experimento um número razoável B de vezes e tentar aproximar a distribuição de $\hat{\theta}$ a uma distribuição obtida empiricamente (EFRON, 1979).

Buscando-se subsídios para a argumentação da metodologia bootstrap, pode-se dizer que, em muitas situações práticas, não se pode realizar esse procedimento (torna-se impraticável) ou por razões econômicas ou devido às condições experimentais que não permitem o princípio da repetição, por causa da própria natureza do fenômeno. Por exemplo, em modelagem espacial (ou predição espacial) utilizando as técnicas de Geoestatística, objeto de estudo desta pesquisa, não há a mínima possibilidade de se replicar o experimento de campo, isto é, a amostra tomada constitui-se de apenas uma única realização. Imagine uma situação muito peculiar em agricultura de precisão, em que se deseja amostrar uma região com o objetivo de se avaliar a taxa de infiltração básica de um solo classificado como Latossolo Vermelho-Amarelo. Trata-se de um sistema de amostragem determinístico, tipicamente em forma de *lattice* (podendo ser regular ou não), cujos indivíduos amostrados são georreferenciados por um sistema de coordenadas arbitrária

- o sistema de coordenadas do espaço euclidiano. Perceba a impossibilidade de se repetir o ensaio porque não se pode fazer uma réplica dessa área de estudo (desse campo aleatório), a qual é uma complexa estrutura espacialmente correlacionada cuja configuração e formação se deram unicamente na gênesis do processo, sendo, portanto, impossível reproduzí-la ou repetir esse evento - o que foi feito, se fez ....

Em suma, toda essa abordagem introdutória da Estatística — exceto no que se refere à Geoestatística, por ser um campo bem específico da Estatística espacial - até aqui relatada é bastante trivial no contexto da Estatística (experimental) e de suma importância para se compreender os procedimentos básicos do bootstrap; portanto, não carece de detalhamento ou exemplos de aplicação. Todavia, se houver a necesidade de maiores esclarecimentos teóricos desse assunto, o próprio leitor deve buscar recursos nas literaturas clássicas de estatística básica ou em livros mais avançados que tratam o assunto inferência e, ainda, em textos de Geoestatística para compreender os fundamentos da amostragem de processos estocásticos correlacionados no espaço. Sugerem-se, para Geoestatística, os autores: Journel (1991); Cressie (1993); Wackernagel (2003); Schabenberger e Gotway (2005); Webster e Oliver (2001) e, para a Estatística convencional, os livros de Cochran(1982); Thompson (2002) podem ser consultados.

2.8.3 Teoria bootstrap: uma ideia genial

Segundo Zoubir e Boashash (1998), o bootstrap é uma poderosa técnica para avaliar a acurácia de um estimador de parâmetro em situações nas quais as técnicas convencionais não são válidas. Nesse artigo eles destacam os principais motivos que o levaram a utilizar os métodos bootstrap na aplicação típica de processos de sinais, mostrando vários exemplos práticos.

Moore (2006) elucida que a reamostragem bootstrap não cria um novo conjunto de dados e, tampouco, as observações reamostradas são tratadas como tal. Apenas se utiliza a distribuição da estatística de interesse das reamostras, para estimar como essa estatística amostral ($\hat{\theta}(z)$) deveria variar, devido à amostragem aleatória. É perfeitamente legítimo se estimar um parâmetro, como também a variabilidade de sua estimativa, usando as técnicas bootstrap. Isto é, exatamente, como se procede na Estatística convencional, quando se deseja calcular $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ para estimar o parâmetro μ e, a partir dessas estatística, calcular o erro padrão da média $\bar{z}$ dado por $s/\sqrt{n}$.

Segundo Rizzo (2008), as estimativas bootstrap de uma distribuição de amostragem são análogas à ideia da estimação de densidade. Assim, constrói-se um histograma de uma amostra para obter uma estimativa do fomato da função de densidade. Sabe-se que o histograma não é a densidade, mas em uma abordagem não paramétrica, este pode ser visto como uma estimativa razoável da densidade. Atualmente, existem métodos para se gerar amostras aleatórias completamente especificadas; o bootstrap é capaz de gerar amostras aleatórias a partir da distribuição empírica da amostra.

Agora, dar-se-á uma definição mais formal do bootstrap de Bradley Efrom (EFRON, 1979). O bootstra é aplicado a uma amostra *i.i.d.* $z' = (z_1, \ldots, z_n)'$ com uma função de densidade de probabilidade $F(z)$. Daí, reamostras *iid* são geradas sorteando-se n vezes com reposição a partir da amostra original $z_1, \ldots, z_n$. Esse procedimento fornece uma reamostra $z_1^*, z_2^*, \ldots, z_n^*$. Em outras palavras, a reamostra é construída gerando $z_1^*, z_2^*, \ldots, z_n^*$ que são condicionalmente independentes (dado o conjunto de dados original) e tem distribuição condicional $\hat{F}(z)$, a qual é denominada distribuição empírica.

Se Z^* é selecionado ao acaso a partir da amostra z, tem-se que

$$P(Z^* = z_i) = \frac{1}{n}, \qquad i = 1, \ldots, n.$$

A partir dessa formulação, o bootstrap, por reamostragem, gera uma amostra aleatória

$$Z_1^*, \ldots, Z_n^* \qquad i = 1, \ldots, n,$$

com reposição, obtida a partir do vetor z. Evidentemente, as variáveis aleatórias Z_i^* são *iid*, uniformente distribuídas sobre o conjunto $\{z_1, \ldots, z_n\}$ (RIZZO, 2008; CHERNICK, 2008). Em outras palavras, a distribuição de probabilidade empírica $\hat{F}$, colocando massa de probabilidade $1/n$ sobre os dados z_i, gera uma amostra aleatória *iid* $z^* = (z_1^*, \ldots, z_n^*)$, isto é, $z_1^*, \ldots, z_n^* \sim \hat{F}$

2.8.4 O princípio *plug-in*

De acordo com Efron e Tibshirani (1993), o princípio *plug-in* é um simples método de estimar parâmetros a partir de amostras. A estimativa *plug-in* de um parâmetro $\theta = \wp(F)$ é definido como

$$\hat{\theta} = \wp(\hat{F}). \tag{2.30}$$

Descrevendo em palavras, estima-se a função $\theta = \wp(F)$ da distribuição de probabilidade de F pela mesma função empírica $\hat{F}$, $\hat{\theta} = \wp(\hat{F})$. Observe que as estatísticas, como definido em 2.30, são usadas para estimar parâmetros, sendo às vezes, denominadas de estatística de resumo, assim como estimativas e estimadores.

Uma interessante definição de Moo (2006) sobre o princípio *plug-in* é a

seguinte: "para estimar um parâmetro, a quantidade que descreve a população, use a estatística que é a correspondente quantidade para a amostra."

Em resumo, o princípio *plug-in* é uma técnica para estimar uma média populacional μ pela média amostral $\bar{x}$ e um desvio padrão da população σ pelo desvio padrão da amostra s. Ainda, estime a mediana populacional pela mediana amostral e estime uma linha de regressão populacional pela linha dos quadrados mínimos a partir da amostra. Perceba que o princípio *plug-in* nada mais é que o velho método dos momentos, tão usual em Estatística. Da mesma forma, a ideia do bootstrap, por si só, é uma forma do princípio *plug-in*: usa o dados (amostra) em vez da população, então, toma reamostras para reproduzir (imitar) o processo de construção de uma distribuição de amostragem.

Segundo Rizzo (2008), para gera uma amostra aleatória bootstrap reamostrando z, geram-se n inteiros aleatórios $\{i_1, \ldots, i_n\}$ uniformemente distribuídos sobre $\{1, \ldots, n\}$ e seleciona-se a amostra bootstrap $\{z^* = (z^*_{i_1}, \ldots, z^*_{i_n})\}$.

Suponha que θ (θ pode ser um vetor) é o parâmetro de interesse e $\hat{\theta}$ é um estimador de θ. Então, o algoritmo bootstrap para se estimar a distribuição de $\hat{\theta}$ é como segue (RIZZO, 2008):

1. Para cada repetição bootstrap, indexada por $b = 1, \ldots, B$:

   (a) gerar a amostra $z^*(b) = (z^*_1, \ldots, z^*_n)$ reamostrando com reposição a partir da amostra observada $z_1, \ldots, z_n$;

   (b) calcular a b-ésima repetição $\hat{\theta}^*(b)$ a partir da b-ésima amostra bootstrap $z^*(b)$.

2. A estimativa bootstrap de $F_{\hat{\theta}}(\cdot)$ é a distribuição empírica das repetição

$$\hat{\theta}^*(1), \hat{\theta}^*(2), \ldots, \hat{\theta}^*(B).$$

O princípio do bootstrap pode ser representado pelo diagrama exibido na Figura 2.19.

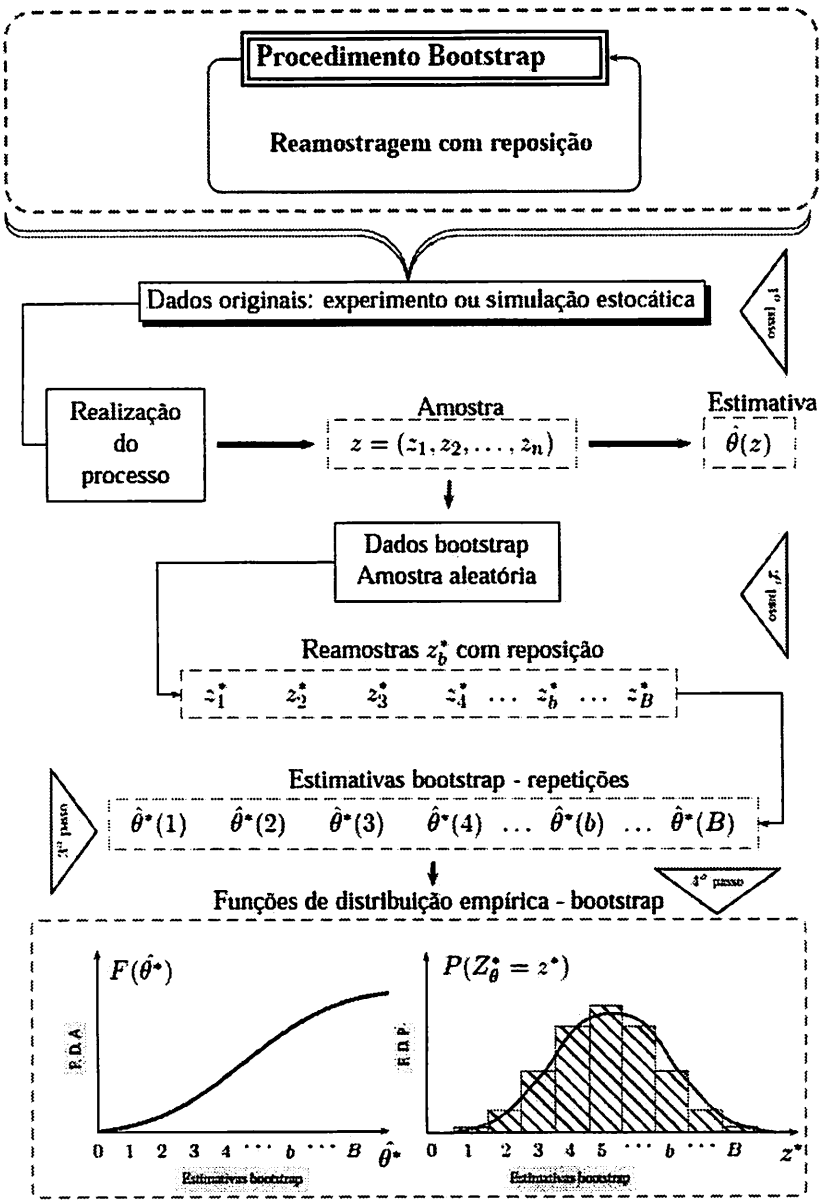


Figura 2.19 Princípio básico de bootstrap

Moore (2006) mostra que a distribuição bootstrap corresponde com a distribuição de amostragem em sua forma e dispersão, mas o centro da distribuição bootstrap não é o mesmo que na distribuição amostral. Note que a distribuição de amostragem de uma dada estatística usada para estimar um parâmetro está centrada no atual valor do parâmetro populacional mais um viés associado. A distribuição bootstrap está centrada no valor da estatística para a amostra original mais um viés. O ponto chave é que os dois vieses são similares, ainda que os dois centros possam não ser. Em vista disso, o uso da técnica bootstrap é mais interesante em situações em que a distribuição de amostragem não é conhecida.

O termo viés (ν) é central na Estatística. Uma estatística é viesada, como estimativa do parâmetro, se sua distribuição de amostragem não está centrada no verdadeiro valor do parâmetro. Colocando de forma mais simples, o viés de uma estatística é calculado pela diferença entre a média de sua distribuição de amostragem e o verdadeiro valor do parâmetro. De modo similar, a estimativa do viés bootstrap é a diferença entre a média da distribuição bootstrap e o valor da estatística na amostra original. Isso pode ser melhor compreendido por meio do desenvolvimento, conforme segue.

Considere algum estimador $\hat{\theta}$ do desconhecido parâmetro θ, em que

$$\hat{\theta} = \wp(z_1, \ldots, z_n).$$

Seja $\hat{\theta}_a$ a estimativa usando os dados originais, enquanto $\hat{\theta}_b$ é a estimativa usando a b-ésima amostra bootstrap. A média de todos os B estimadores bootstrap é dada por

$$\hat{\theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b. \tag{2.31}$$

Por meio de $\hat{\theta}^*$, obtém-se uma estimativa do viés ν do estimador $\hat{\theta}$:

$$\nu = E[\hat{\theta}] - \theta, \tag{2.32}$$

em que $\hat{\theta}^*$ é a estimativa bootstrap de $E[\hat{\theta}]$. Então, pode-se escrever, $\hat{\nu} = \hat{\theta}^* - \hat{\theta}_a$, esperando-se uma estimativa de viés bootstrap $\hat{\nu} = 0$, se e somente se:

$$E[\hat{\theta}^*] - \hat{\theta}_a = 0 \tag{2.33}$$

$$E[\hat{\theta}^*] = \hat{\theta}_a \tag{2.34}$$

2.8.5 Estimativa do erro padrão bootstrap

Outra importante medida para os estatísticos é a estimativa do errro padrão. Suponha que foram geradas as amostras bootstrap $z_1^*, \ldots, z_n^*$. Se $\hat{\theta}$ é a média amostral, por exemplo, então, $\hat{\theta}^*(b)$ é a média da b-ésima amostra bootstrap, tal que $b = 1, \ldots, B$. A estimativa do erro padrão bootstrap é o desvio padrão das repetições bootstrap (Efron, 1993),

$$\widehat{SD}_b = \sqrt{\sum_{b=1}^{B} \frac{(\hat{\theta}^*(b) - M_b)^2}{B-1}} \qquad M_b = \sum_{b=1}^{B} \frac{\hat{\theta}^*(b)}{B}, \tag{2.35}$$

Segundo Efron e Tibshirani (1993), o número de repetições B para estimar um erro padão não precisa ser muito grande; $B = 50$ é o suficiente, e raramente se utiliza um valor $B > 200$. Porém, para a estimação de intervalos de confiança, o valor de B deve ser muito maior.

Na Figura 2.20 apresenta-se um diagrama ilustrando o procedimento para o cálculo da estimativa do erro padrão bootstrap.

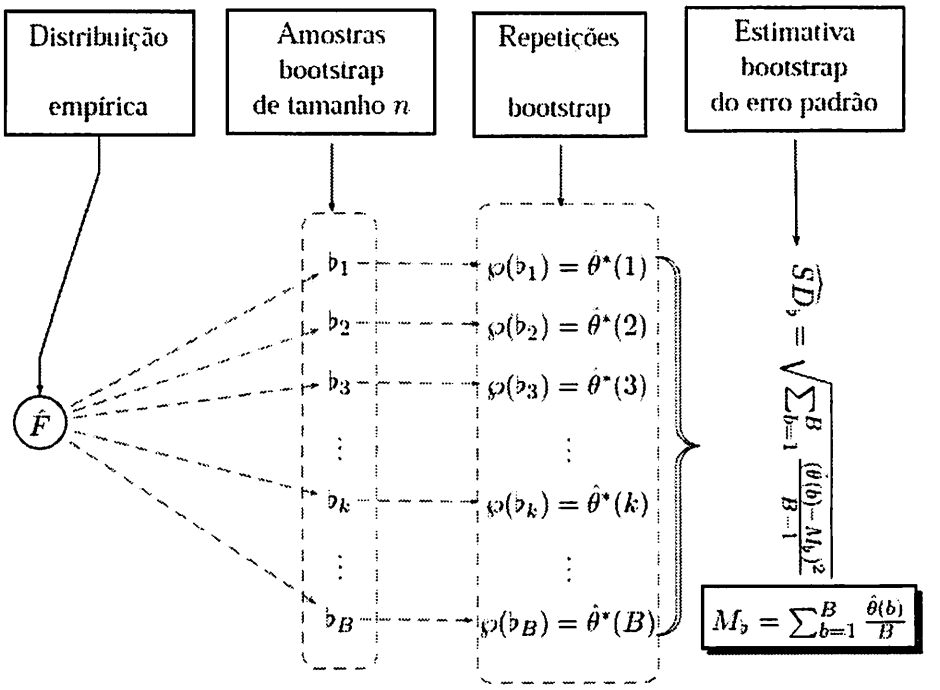


Figura 2.20 Algoritmo bootstrap para estimar do erro padrão, $\widehat{SD}_b$, de uma estatística $\hat{\theta}$; cada amostra bootstrap é uma amostra aleatória independente de tamanho n gerada a partir da função de distribuição empírica $\hat{F}$

Fonte: Adaptação de Efron e Tibshirani (1993)

2.9 Limites de confiança percentil bootstrap

O intervalo de confiança percentil bootstrap (ICPB) é construído com base na distribuição das repetições bootstrap. A ideia é usar os percentis do bootstrap (histograma) para definir os limites de confiança. Assim, os quantis da distribuição empírica são os estimadores dos quantis da distribuição de amostragem de $\hat{\theta}$ e esses quantis podem melhor se ajustar à verdadeira distribuição quando as distribuição de $\hat{\theta}$ não segue uma distribuição normal. Na ausência de normalidade nos dados, o método percentil tem algumas vantagens teórica sobre o intervalo da normal

padrão, apresentando um melhor desempenho que este. Sejam $\{\hat{\theta}^*(1), \ldots, \hat{\theta}^*(B)\}$ as repetições de $\hat{\theta}$. Por meio da função de distribuição acumulada empírica das reamostras bootstrap, o ICPB é construído calculando-se o $\alpha/2$ quantil $\hat{\theta}^*_{\alpha/2}$, e o $(1-\alpha/2)$ quantil $\hat{\theta}^*_{1-\alpha/2}$.

Efron e Tibshirani (1993) usaram um pequeno ensaio clínico com ratos, a fim de avaliar o desempenho do ICPB, em comparação com o intervalo da normal padrão. Trata-se de uma amostra de tamanho $n = 16$, apresentando uma distribuição fortemente assimétrica, isto é, os dados não possuem normalidade. Após $B = 1000$ repetições, os percentis de $\hat{\theta}^*$ foram calculados e o ICPB de 95% foi construído. Esse estudo mostra que ICPB foi mais preciso que o intervalo da normal padrão e, somente após aplicar uma tranformação logarítmica nos dados, ambos os intervalos tiveram desempenho similar. Esses resultados formalizam o fato de que o método percentil pode ser pensado como um algoritmo para automaticamente incorporar tais transformações, isto é, o intervalo percentil para θ se harmoniza bem com o intervalo de confiança da normal padrão sobre um transformação apropriada.

A metodologia bootstrap permite construir diversos tipos de intervalos de confiança, a saber: o intervalo-t bootstrap, intervalo BCa (*bias-corrected and accelerated*), ABC e ABCq. Porém, há interesse em apresentar apenas o intervalo percentil e o leitor interessado nos outros intervalos de confiança, devem consultar Efron e Tibshirani (1993); Chernick (2008); Rizzo (2008); MANLY (1991).

2.10 Sumário

Inspirado por Moore (2006), apresenta-se um pequeno resumo sobre os fundamentos básicos da metodologia bootstrap.

- Good (2005) decreveu que a característica mais atrativa do bootstrap é a li-

berdade que ele fornece às hipóteses paramétricas restritivas e simplificados. Assim, não há necessidade de forçar normalidade ou quaisquer outras suposições distribucionais paramétricas sobre os dados. Em muitos problemas, os dados podem ser assimétricos ou ter distribuição de caudas pesadas ou ainda ser multimodal. O modelo não necessita ser simplificado para alguma aproximação linear e o estimador por si mesmo pode ser complicado.

- Para realizar o bootstrap em uma estatística como a média amostral, tome centenas de reamostragem com reposição a partir da amostra original, calcule a estatística para cada reamostra e inspecione (avalie) a distribuição bootstrap das estatísticas reamostradas.

- Uma distribuição bootstrap, geralmente, tem, aproximadamente, o mesmo tipo e dispersão que a distribuição da amostragem.

- Utilize sumários numéricos e gráficos para determinar se a distribuição bootstrap é aproximadamente normal e centrada na estatística original, e para se ter uma ideia de sua dispersão — o erro padrão bootstrap é o desvio padrão da distribuição bootstrap.

- O bootstrap não substitui nem aumenta os dados originais. Usa-se sua distribuição como uma maneira de estimar a variação em uma estatística baseado nos dados originais.

- Conforme descrito em Efron e Tibshirani (1993, p. 173): para dados que apresentam assimetria, o método percentil bootstrap sempre conhece a transformação correta, automaticamente — o método percentil pode ser pensado como um "logaritmo" para incorporar tais trasformações.

- Segundo Chernick (2008), outra característica que faz a abordagem boots-

trap atrativa é sua simplicidade. Podem-se formular simulações bootstrap para quase todo problema concebível. Uma vez programado o computador para executar as repetições bootstrap, deixa-se todo o trabalho para o computador. Um perigo para esta abordagem é que um usuário pode usar os métodos bootstrap sem consultar um estatístico (ou desconsiderar as implicações estatatísticas) e sem refletir cuidadosamente sobre o problema.

2.11 Notas históricas

Michael R. Chernick realizou uma interessante pesquisa abordando toda a evolução teórica do método bootstrap, apresentada em Chernick (2008) e alguns tópicos desses relatos serão reportados a seguir, no principal intuito de mostrar o amplo poder de aplicação desse método para resolver variados poblemas de pesquisa, em quaisquer área de estudo, de maneira bastante simples. Além disso, com essas informações, deseja-se, ainda, tentar esclarecer possíveis dúvidas ou questionamentos sobre a eficiência do bootstrap, relatando, cronologicamente, uma gama de exemplos aplicados e pesquisas bem sucedidas que comprovam ser o método bootstrap uma ferramenta de análise estatística já consolidada e muito útil.

Deve-se salientar que, embora as investigações sobre bootstrap se iniciaram no final de 1979, muitos desenvolvimentos importantes se deram antes dessa data. Mas, as primeiras provas da consistência da estimativa bootstrap da média amostral vieram em 1981, com os artigos de Seingh (1981); Bickel e Freedman (1981). Esses autores apresentaram os primeiros resultados demonstrando a consistência do bootstrap sob certas condições matemáticas.

Segundo Davison e Hinkley (1997), o artigo inaugural de Efron (1979) — o primeiro artigo publicado em 1979 de Bradley Efron sobre os métodos bootstraps — foi um evento muito importante em Estatística que, ao mesmo tempo, sintetizou

algumas das ideias de reamostragem já existentes e estabeleceu uma nova estrutura para análise estatística baseada em simulação. De fato, isso foi um pensamento (um *insight*) revolucionário: a brilhante perspicácia de substituir aproximações imprecisas e complicadas para vieses, variâncias e outras medidas de incertezas, por simulações de computador, chamou a atenção de ambos pesquisadores teóricos e a de usuários de métodos estatísticos de modo geral.

Lehmann e Casela (1999) afirmam que o bootstrap é uma ferramenta que reduz o viés do estimador alcançado, obtendo-se estimativas mais eficientes. Lehmann (1990) apresenta alguns detalhes sobre as propriedades assintóticas do bootstrap.

Os métodos de reamostragem começou com Hartigan (1969,1971,1975) e McCarthy (1969) e, depois desse autores, um estudo sobre esse assunto pode ser verificado em Babu (1992). E, quanto à estimação de quantis via bootstrap, surgiram em Helmers, Janssen e Veraverbeke (1992) e em Falk e Kaufmann (1991).

A terminologia bootstrap tem sido usada em outros contextos semelhantes aos que antecedem o trabalho de Efron (1979). Porém, esses métodos não são os mesmos e, portanto, podem gerar alguma confusão na mente do leitor. Por esse motivo, Chernick (2008) adverte sobre a existência de outros procedimentos diferentes, mas com uma ideia similar àquela de Efron. Para certificação do leitor, Chernick (2008) mencionou que, ao realizar uma apresentação sobre o bootstrap no ano de 1983, o Dr. Ira Weiss (membro da *Corporation Aerospace College*; o autor não menciona a localidade do ocorrido), o informou que ele usou o "bootstrap" em 1970, muito antes de Bradley Efron cunhar o termo. Depois disso, Chernick (2008) avaliou o artigo de Weiss (1970) e percebeu que se tratava de uma abordagem similar, mas, com procedimentos distintos. Sabe-se, ainda, que alguns pequisadores lançaram mão de um procedimento para aplicar o fitro de *Kalman*

com uma estrutura de covariância de "perturbação" ou "ruído" desconhecido, o qual eles também denominaram de bootstrap. Certamente, todos esses *estudiosos*, cada um em suas respectivas épocas, tiveram, igualmente, a mesma inspiração, baseados na ficção em que o Barão Von Munchausen se utiliza de um truque em que ele próprio se retira do fundo de um lago, e daí veio a expressão: "picking yourself up by your own bootstraps" (CRAWLEY, 2006). De fato, parece fazer sentido o uso do termo bootstrap porque o mesmo se aplica a um processo de estimação que evita uma suposição a priori e somente usa os dados diponíveis em mãos. O termo bootstrap também tem sido utilizado em contextos totalmente diferentes por cientistas da computação (CHERNICK, 2008). A propósito, por causa das variantes citadas acima, tem-se o cuidado de esclarecer que a metodologia bootstrap abordada neste trabalho está totalmente fundamentada na teoria formalizada em Efron (1979, 1993) e essa preocupação se justifica pelo fato de que muito poucos autores fazem essa distinção.

Somente para ratificar a importância dessa informação, na literatura estatística quase sempre há referências ao bootstrap de Efron ou a algumas derivações dele. Porém, em literatura para engenharias pode existir uma ambiguidade e, portanto, recomenda-se fortemente avaliar, cautelosamente, a natureza do procedimento que está sendo descrito, a fim de discriminar com precisão qual a abordagem adotada pelo autor (CHERNICK, 2008).

3 MATERIAIS E MÉTODOS

3.1 Informações a priori

A metodologia foi totalmente desenvolvida com fins didáticos, computacionais e estatísticos. Dessa forma, intenta-se facilitar o aprendizado e, ao mesmo tempo, integrar os métodos abordados a uma apropriada linguagem de computação, disponíveis no ambiente do program R.

Incrementaram-se ou incorporaram-se nos procedimentos analíticos da modelagem do semivariograma/variograma os benefícios das técnicas de reamostragem via método bootstrap. Nesse contexto, grandes foram as vantagens que essa abordagem pode oferecer àqueles que se utilizam da modelagem variográfica para fazer predições de processos estocásticos espaciais de variáveis contínuas, principalmente. Tais inovações possibilitaram a construção de valiosos intervalos de confiança para os parâmetros de covariância espacial (efeito pepita, patamar e alcance) e viabilizaram, também, a elaboração de semivariogramas de quantis, dando ao pesquisador mais poder de análise. Isto é, tornou-se possível avaliar as incertezas associadas na estimação do semivariograma experimental ($\hat{\gamma}(h_k)$), como também no seu modelo teórico ajustado ($\gamma(h_k;\theta)$)

Além disso, o uso do bootstrap, notadamente, suscitou um teste de normalidade multivariada para Geoestatística baseada em modelos. A ideia central é usar o semivariograma experimental como base para gerar uma estatística de teste e o suporte teórico para isso está relacionado com dois dos métodos usuais de ajustes do semivariograma teórico: trata-se do método dos quadrados mínimos que é ajustado diretamente sobre a nuvem de pontos do semivariograma experimental e o método da máxima verossimilhança, que ajusta o modelo com base nos dados amostrados, ignorando completamento o gráfico de semivariâncias. O bootstrap

foi essencial na construção de intervalos de confiança para avaliar o comportamento desses métodos, em presença de normalidade multivariada. Dessa forma, elaborou-se uma estatística de teste capaz de quantificar a diferença significativa entre aqueles ajustes (valor de plausibilidade, *valor-pl*) e, portanto, aceitar ou rejeitar a hipótese H_0, previamente postulada, de os dados espaciais possuírem distribuição normal multivariada. Essa metodologia é, basicamente, uma ferramenta gráfica apropriada para diagnosticar o comportamento da massa de dados e checar as evidências que ratifiquem, ou não, a existência de normalidade multivariada entre os indivíduos amostrais.

Todos os métodos de análises propostos aqui foram executados no ambiente R, utilizando os recursos computacionais disponíveis nos pacotes RandomFields (SCHLATHER, 2010) e geoR (RIBEIRO JUNIOR; DIGGLE, 2001), especialmente desenvolvidos para se realizar análise geoestatística com opções para simulação de campos aleatórios de modo geral. Utilizaram-se também rotinas (funções) computacionais desenvolvidas especificamente para essa tese. Outro aspecto computacional importante foi a facilidade de se obter ferramentas de otimizações não lineares já implementadas nas funções optim, optimize, nlm e nls, nls2, pertencentes ao pacote stats (R DEVELOPMENT CORE TEAM, 2009), as quais foram de grande valor neste trabalho. Assim, por meio desse conjunto de funções, pode-se facilmente utilizar e avaliar a metodologia proposta.

Um detalhe nesse trabalho é a sugestão do uso da razão entre o número de pares de pontos $N(h)$ e qualquer potência do inverso da distância h ($W_{N/h} = \frac{N(h)}{h^p}$: p = 1,2,3,...) como ponderação para o método dos quadrados mínimos ponderado (QMP). Trata-se de um interessante critério de peso que se justifica pela lógica da própria natureza dos fenômenos espaciais. Em outras palavras, esse critério valoriza a essência da Geoestatística: a razão da distância (h) leva em conta

o fato de os dados serem uma estrutura regionalizada e, portanto, o fator de ponderação diminui em função da distância, comportando-se semelhante à covariância, posto que o $\lim_{h \to \infty} C(h) = 0$. Esse fato torna o referido critério de ponderação, além de fácil implementação computacional, bastante coerente com a teoria das variáveis regionalizadas. Nas próximas seções descreve-se detalhadamente cada um dos procedimentos ou métodos expostos na Seção 3.1.

3.2 Bootstrapping: um exemplo de simulação

Antes de apresentar a metodologia propriamente dita, convém introduzir um simples exemplo usando dados simulados para mostrar como se processa o método bootstrap na Estatística convencional e, concomitantemente, mostrar sua implementação computacional no programa R, bem como a obtenção de algumas estatísticas elementares que o mesmo possibilita.

Exemplo didático 1

Seja uma amostra x com distribuição $N(3; (0{,}25)^2)$, de tamanho $n = 10$, gerada pela função *rnorm* do programa R [4]:

Nota: O camando `set.seed(.)`, denominado semente, permite ao leitor reproduzir os mesmos resultados aqui apresentados.

Programa1. Primeiramente, gera-se, facilmente, um vetor de dados, x_{10}, executando a função `rnorm`, conforme demonstrado a seguir:

```
set.seed(1) # semente para reprodução
 # gera o vetor de dados normais
```

[4] Assume-se que leitor já tenha um conhecimento básico do programa R para entender os códigos e a forma como o R retorna os resultados.

```
> x <- rnorm(n=10,mean=3.0,sd=0.25)
> print(x)
[1] 2.843387 3.045911 2.791093 3.398820 3.082377
[6] 2.794883 3.121857 3.184581 3.143945 2.923653
# média e desvio padrão de x
> c(x.bar=mean(x),dp=sqrt(var(x)))
x.bar      dp
3.0330507 0.1951465
```

Programa2. A seguir, com a função `sample` (também pode-se usar os pacotes: `boot` e `bootstrap`), aplica-se a técnica de reamostragem com reposição (bootstrapping) a partir do vetor amostral x de dimensão 10. Isso significa que, com equiprobabilidade $\frac{1}{10}$, cada valor $\{x_i \in x\}$ pode ser tomado mais de uma vez e pode ocorrer de um ou mais valores não comporem a *b-ésima* amostra do conjunto das B repetições bootstrap (reamostragem) $\{x_b^* : b = 1,\ldots,B\}$. Veja o resultado de 3 reamostras selecionadas por essa técnica, via função `sample` e a frequência com que seus elementos aparecem em cada amostra bootstrap, x_1^*, x_2^* e x_3^*.

```
set.seed(11)
# Primeira repetição bootstrap
> x.ast1 <- sample(x,replace=TRUE)
set.seed(12)
# Segunda repetição bootstrap
> x.ast2 <- sample(x,replace=TRUE)
set.seed(122)
# Terceira repetição bootstrap
> x.ast3 <- sample(x,replace=TRUE)
```

```
# tabelas de frequências das repetições bootstrap
> print(list(x.ast1=table(round(x.ast1,3)),
+ x.ast2=table(round(x.ast2,3)),
+ x.ast3=table(round(x.ast3,3))))
$x.ast1
 2.791  2.795  2.843  2.924  3.046  3.144
     2      1      4      1      1      1
$x.ast2
 2.791  2.843  2.924  3.046  3.122  3.144
     1      4      1      2      1      1
$x.ast3
 2.795  2.843  2.924  3.046  3.122
     2      3      2      1      2
```

Comentário1. Observando-se as resultados emitidos acima, x.ast1, x.ast2 e x.ast3 são as amostras bootstrap tomadas aleatoriamente do vetor x. O dado 2,843 ocorreu 4 vezes nas duas primeiras retiradas e 3 vezes na terceira, enquanto o dado 3,144 foi sorteado apenas uma vez, nas duas primeiras amostras e nenhuma vez na última repetição bootstrap efetuada. De forma semelhante ocorreu com os demais elementos nas reamostras de modo geral.

Programa 3. Sabe-se que o método boostrap estima a distribuição de uma população por reamostragem. Assim, a amostra (x) observada é considerada uma "população" finita (pseudopopulação) de onde muitas amostras aleatórias são geradas (repetições bootstrap) e, então, qualquer procedimento inferencial pode ser aplicado para estimar as características de interesse da real população com base na distribuição empírica obtida por meio das repetições bootstrap. Essencialmente, a média, o desvio padrão e os intervalos de confiança são importantes estatísticas a serem estimadas. Assim, será apre-

sentado a seguir como gerar a distribuição empírica bootstrap e construir o intervalo de confiança com 95% de confiabilidade para a média populacional.

```
# preparando o bootstrap:
> B = 999     # número de repetições
> n = length(x) # tamanho da amostra
> teta.ast <- numeric(B) # armazenar as repetições
# aplicando o bootstrap:
> for (b in 1:B){
# indexa as repetições
+ i  <- mean(sample(1:n,size=n,replace=TRUE))
+ teta.ast[b] <- mean(x[i])}
> utils::str(teta.ast)
num [1:999] 3.07 3.05 3.03 2.99 3.07 ...
> media.boots <- mean(teta.ast)
> dp.boots <- sqrt((1/(B-1))*
+ (sum((teta.ast-media.boots)^2)))
# média e desvio padrão da distribuição bootstrap
> print(c(teta.media=media.boots,teta.dp=dp.boots))
# viés bootstrap
> media.boots-est.x[[1]]
[1] -0.00123455
```

Comentário2. A rotina computacional do programa 2 executa a técnica bootstrap sorteando as 999 amostras $\{x_1^*, x_2^*, \ldots, x_B^*\}$ e, simultaneamente, calcula suas respectivas médias $\{\theta_1^*, \theta_1^*, \ldots, \theta_B^*\}$. Assim, gera-se a distribuição empírica das médias amostrais e, consequentemente, obtêm-se os quantis

2,5% e 97,5% (próximo item), com os quais se pode construir o intervalo de confiança percentil bootstrap. Outros tipos de intervalos de confiança bootstrap poderiam ser calculados como aqueles descritos em Efron e Tibshirani (1993).

Programa 4. Segue o procedimento para se calcular o intervalo de confiança percentil bootstrap e a construção do histograma da distribuição empírica das médias $\{\theta_1^*, \theta_1^*, \ldots, \theta_B^*\}$ reamostradas do vetor x. O histograma está disponível na Figura 3.1.

```
 hist(teta.ast,border="black",plot=T,
+ xlab="Teta.asterisco",ylab="Frequência",
+ main="",ylim=c(0,350))
points(mean(teta.ast),0,col=1,pch=4,cex=1)
segments(mean(teta.ast),0,mean(teta.ast),324,
+ lty=2,lwd=1.5)
text(mean(teta.ast),326,paste(round(media.boots,3)),
+ cex=.8,pos=4,offset=0)
(quantil <- quantile(teta.ast,c(0.025,0.975)))
+ 2.5%    97.5%
2.926102 3.153046
segments(quantil[[1]],0,quantil[[1]],324,
+ lty=2,lwd=1.5)
points(quantil[[1]],0,col=1,pch=4,cex=1)
text(quantil[[1]],326,
paste(round(quantil[[1]],3),"(2,5%)"),
+ cex=.8,pos=4,offset=0)
segments(quantil[[2]],0,quantil[[2]],324,
```

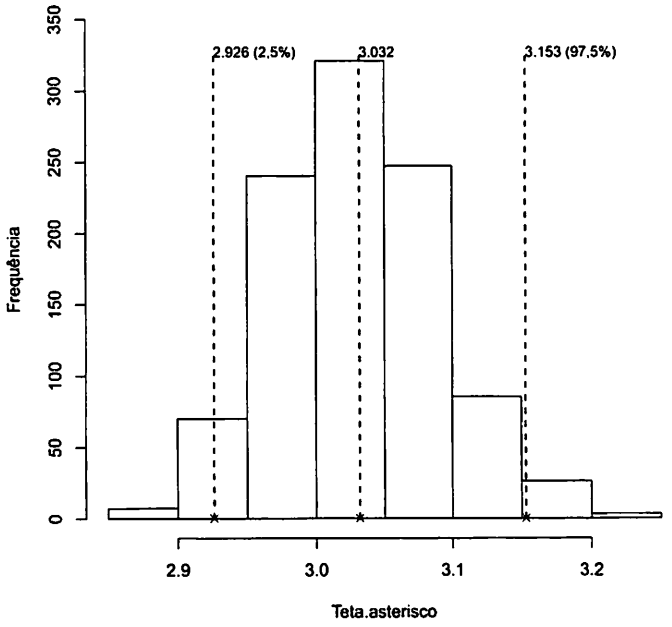


Figura 3.1 Histograma da distribuição empírica bootstrap

```
+ lty=2,lwd=1.5)
points(quantil[[2]],0,col=1,pch=4,cex=1)
text(quantil[[2]],326,
+ paste(round(quantil[[2]],3),"(97,5%)"),
+ cex=.8,pos=4,offset=0)
```

Obviamente, a abordagem apresentada anteriormente, tão usual na estatística tradicional, torna a metodologia bootstrap um grande atrativo para se fazer inferência sobre as incertezas associadas aos princípios da modelagem geoestatística.

3.2.1 Aspectos computacionais gerais e simulação estocástica

Utilizou-se a função `variog` do pacote geoR (RIBEIRO JUNIOR, 2001) para se estimar o semivariograma de nuvens e construir a base de dados ($\Gamma(h)$) para aplicação da metodologia bootstrap, como também a função `likfit` apenas para a modelagem variográfica pelo critério da MV. Já toda a modelagem do semivariograma efetuada pelos métodos dos QM foi executada por meio das funções de correlações espaciais construídas pelo próprio autor: `SphFit`, `GauFit`, `ExpFit`. Essas funções processam os ajustes não lineares com o auxílio da função de minimização `nlm`, que usa um algoritmo tipo-Newton, e da função de otimização `optim`, ambas implementadas no pacote `stats` do programa *R* (DEVELOPMENT CORE TEAM, 2009). No momento, é pertinente esclarecer que o teste de normalidade multivariada para Geoestatística disponibiliza ferramentas para avaliar somente as estruturas de correlação espacial esférica, gaussiana e exponencial, mas, com projeto futuro de alcançar todos os modelos de Geoestatística (em torno de 43 modelos) disponíveis atualmente na literatura.

A estratégia para avaliar o desempenho desse teste foi criteriosamente planejada, combinando-as em três cenários. No primeiro cenário, o conjunto de dados utilizados é proveniente de simulações que representam realizações de processos estocásticos univariados gaussianos, assumindo, em ambos os casos, variáveis estacionárias, sem tendência direcional, isto é, isotropia. No segundo cenário, simularam-se processos não gaussianos mantendo as mesmas demais suposições adotadas no primeiro cenário. Detalhando, nesse ambiente foram simulados três processos gaussianos e três processos não gaussianos e essas simulações estocásticas foram realizadas por meio da função *gaussian random fields* (`gaussRF`) implementada no pacote `RandomFields`, desenvolvido por Schlather (2010). Quanto ao terceiro cenário, apreciou-se o referido teste mediante três conjuntos de

dados reais, abrangendo diversas áreas de aplicação. Desse modo, o teste proposto foi avaliado em nove situações diferentes, as quais serão descritas a seguir.

O procedimento para avaliar a performance do teste de normalidade para Geoestatística será explicado em detalhes a seguir. Antes, deve-se esclarecer que, para a construção da estrutura de correlação espacial dos campos aleatórios simulados, optou-se apenas pelos modelos de correlação esférica, exponencial e gaussiana. Essa escolha deve-se, principalmente, ao fato de essas estruturas serem as mais comumente utilizadas na prática. Além desse motivo, em especial, o modelo esférico foi selecionado porque também possui uma estrutura de dependência espacial bem comportada, com um patamar bem definido. Uma outra razão para se realizar a avaliação do teste proposto somente nessas três estruturas de correlação foi o limitador tempo. Porém, em avanços futuros, pretende-se abordar os demais modelos de covariânica espacial autorizados em Geoestatística.

Cenário I: simulação de processos estocásticos gaussianos

- **Modelo com correlação espacial esférica:** configuraram-se 121 pontos sobre uma malha regular de tamanho 11×11, sendo os vetores de coordenadas $x = (0,\ldots,5)$ e $y = (0,\ldots,5)$, tendo um incremento de 0,5 unidades. Dado o modelo esférico, $\gamma(h,\theta) = \tau^2 + \varrho^2\left(1{,}5\frac{h}{\varphi} - 0{,}5\left(\frac{h}{\varphi}\right)^3\right)$, um campo aleatório gaussiano foi gerado, com média 3,0, pelo particular modelo de covariância espacial

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ 0{,}5 + 4{,}5\left(\frac{3}{2}\left(\frac{h}{2}\right) - \frac{1}{2}\left(\frac{h}{2}\right)^3\right) & 0 < h \leq 2 \\ 5{,}0 & h > 2 \end{cases}$$

sendo os seus parâmetros pepita, contribuição e alcance, respectivamente,

iguais a $\tau^2 = 0{,}5$, $\varrho^2 = 4{,}5$ e $\varphi = 2{,}0$

- **Modelo com correlação espacial exponencial:** os dados georreferenciados seguem um padrão espacial exponencial de variável gaussiana e foram simulados sobre um *lattice* regular de dimensão 15 × 15, com limites mínimo e máximo nas coordenadas x e y iguais a 0 e 9,8, $n = 225$ pontos e 0,7 de espaçamento; o conjunto de parâmetros usados no modelo foi $\{\mu = 0; \tau^2 = 0{,}7; \varrho^2 = 6{,}0; \varphi = 3{,}5\}$. Então, o modelo gerador desse campo aleatório gaussiano que será avaliado por esse teste de normalidade é definido como

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ 0{,}7 + 6{,}0\left(1 - e^{\frac{-3h}{3{,}5}}\right) & h \neq 0 \end{cases},$$

em que o processo é estacionário e onidirecional.

- **Modelo com correlação espacial gaussiana:** criou-se uma malha regular de dimensão 10 × 10, com $n = 100$ realizações estocásticas Gaussianas, configurada com espaçamento 0,7, sendo as coordenadas (x,y) mínima e máxima iguais 0 e 6,3. Assim, o padrão de dependência espacial foi simulado sobre o *lattice* regular, segundo o modelo gaussiano

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ 1{,}2 + 10{,}0\left(1 - e^{-3\left(\frac{h}{2{,}6}\right)^2}\right) & h \neq 0 \end{cases},$$

considerando os argumentos $\{\mu = 4.2; \tau^2 = 1{,}2; \varrho^2 = 10{,}0; \varphi = 2{,}6\}$.

Cenário II: Simulação de processos estocásticos não guassianos

Considerando também o modelo com correlação espacial esférica, expo-

nencial e gaussiano, os processos estocásticos foram gerados sobre um *lattice* regular, 11 × 11, com um espaçamento de 0,5 unidades, tendo os limites mínimo e máximo nas coordenadas x e y iguais a 0 e 5. O modelo adotado para gerar os dados não gaussianos relativo a cada um dos modelos de correlação espacial em questão obedece ao seguinte princípio: primeiramente, simulou-se um processo estocástico gaussiano definido pelos argumentos $\{n = 121; \mu = 6{,}0; \tau^2 = 1{,}0; \varrho^2 = 9{,}0; \varphi = 3{,}0\}$ e, depois, esse processo gaussiano foi perturbado com uma função de distribuição exponencial com parâmetro escala (média) $b = 1/3$, descrita no programa R como `rexp(n=121,rate=3)`, fazendo com que a distribuição dos dados seja desconhecida. Desse modo, o teste pôde ser avaliado na ausência de normalidade multivariada nos dados arbitrariamente simulados.

Cenário III: Aplicações em dados reais

- **Área de solos: dados de argila**

  O estudo foi conduzido em lavouras de grãos sob condicões de sequeiro, na fazenda Alto Alegre, na cidade de Planaltina de Goiás, GO, num talhão com sucessão milho-soja, sem preparo do solo há cinco anos. A área apresenta Latossolo Vermelho-Amarelo, com teor de argila igual a 543 g kg^{-1}. A área amostrada para análise espacial pode ser vista na Figura 3.2.

- **Área de solos: dados do rio *Meuse***

  Neste conjunto de dados, a localização (Figura 3.3) e as concentracões de metais pesados das superfície do solo, em ppm, juntamente com variáveis da paisagem e do solo, foram coletados nas planícies aluviais do rio *Meuse*, oeste da povoação de *Stein*, Países Baixos, *Utrecht University*, 1993. As concentrações de metais pesados são magnitudes amostradas de uma área

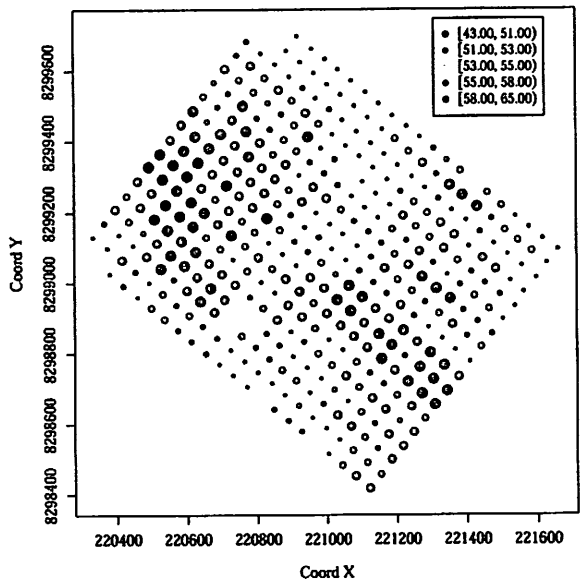


Figura 3.2 *Lattice* regular exibindo os locais onde as amostras de argila foram coletadas
Fonte: os dados foram cedidos por Sandro Manoel C. Hurtado (HURTADO, 2008)

de aproximadamente 15 × 15. Todo o conjunto de dados consiste de 12 variáveis, tomadas em 155 pontos amostrados associados às coordenadas, mas, o teste foi submetido apenas nas variáveis matéria orgânica (MA). Esse conjunto de dados está disponível no pacote `gstat`. Também, utilizou-se a variável cobre, pertencente ao conjunto de dados do rio *Meuse*, para ilustrar o semivariograma de quantis bootstrap.

- **Área de agricultura de precisão: zoneamento de café**

  Os dados sobre produção de café no Brasil (*Coffea arabica* L. e *Coffea sp.*), em toneladas, foram referentes ao período de 1990 a 2005. Contudo, avaliaram-se nesta tese, apenas os dados coletados no ano de 2005. A pro-

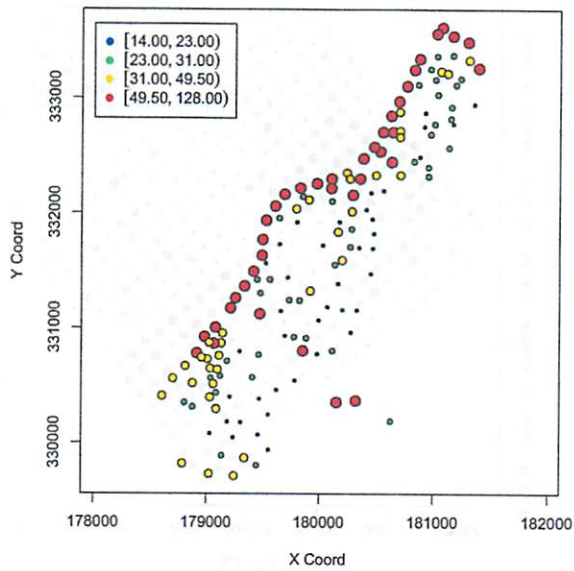


Figura 3.3 Localizações dos pontos amostrados referente aos dados coletados nas planícies do rio *Meuse*, próximo da povoação de Stein, Países Baixo — Universidade de Utrecht, 1993
Fonte: os dados foram obtidos do pacote *gstat* (PEBESMA, 2004)

dução foi obtida pela rede de coleta do Instituto Brasileiro de Geografia e Estatística (IBGE), mediante consulta a entidades públicas e privadas, a produtores, técnicos e órgãos ligados direta ou indiretamente aos setores da produção, comercialização, industrialização e fiscalização de produtos agrícolas. A unidade de investigação no inquérito estatístico foi em município do estado de Minas Gerais, conforme mostrado na Figura 3.4.

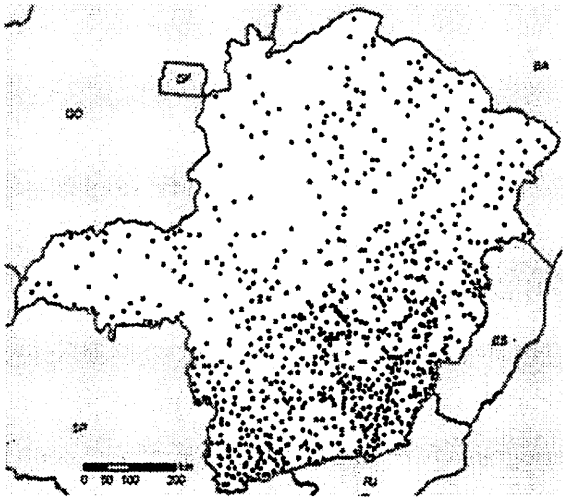


Figura 3.4 Localização das observações nas 853 sedes municipais (■) de Minas Gerais

4 RESULTADOS E DISCUSSÕES

4.1 Resultados metodológicos

Nesta seção, desenvolveram-se os procedimentos da metodologia proposta por esta tese. Em suas particularidades, será abordada a construção do teste de normalidade multivariada em Geoestatística, o qual recebeu o nome de de **Teste Gaussiano para Geoestatística** (TGGeo), por simplificação e, também, apresentar-se-ão os fundamentos da construção do semivariograma de quantis. O TGGeo é o principal objeto deste estudo científico, sendo o semivariograma de quantis um valioso subproduto que foi elaborado (sugiu como uma consequência metodológica subentendida na contrução do TGGeo, obtido, originalmente, por meio da distribuição empírica das semivariâncias bootstrap), paralelamente, no decorrer do desenvolvimento da metodologia central. Esses instrumentos desenvolvidos para análise em Geoestatística tiveram como suporte teórico as técnicas de reamostragem bootstrap. Será apresentado, ainda, um sistema de ponderação, $W_{N/h}$, para modelar o semivariograma experimental (ajuste de curvas) por meio do critério de quadrados mínimos ponderados.

4.1.1 Trabalhando os dados: definições e nomenclaturas

O objetivo deste item é descrever como os dados do semivariograma experimental de nuvens são apurados e preparados, formalmente, para se utilizar os procedimentos bootstrap e obter a distribuição empírica das dissimilaridade em cada classe de distância ou estratos h. Considere um conjunto de variáveis aleatórias univariadas $\{Z(x_i) : i = 1, \ldots, n\}$, medidas sobre um conjunto de pontos x_i pertencentes a um campo aleatório $\mathcal{R}$ bidimensional, com distribuição desco-

nhecida. Assume-se que os dois primeiros momentos de $Z(x_i)$ não dependem de x_i, isto é, o fenômeno possui estacionaridade de primeira e segunda ordem. Seja também $\{h_k : k = 1, \ldots, m\}$ a medida de incremento espacial entre dois pontos x_i e x_j que comutam a mesma distância (mesmo *incremento*); k é um indexador que seleciona os pares de pontos equidistantes, organizando-os em m estratos formados em função da distância $(h_1, h_2, \ldots, h_k)$. A dimensão (dim) de cada estrado é dada por $\{dim(h_1) = n_1, dim(h_2) = n_2, \ldots, dim(h_k) = n_k\}$. De fato, o total de pares N de um dado processo pode ser calculado pela operação aritmética: $N = n_1 + n_2 + \ldots + n_k$. Assim, estimam-se as medidas de dissimilaridades espaciais pela expressão (WACKERNAGEL, 2003),

$$\hat{\gamma}_q(h_k) = \frac{[z(x_i + h_k) - z(x_i)]^2}{2}; \quad k = 1, 2, \ldots, m \quad \text{e} \quad i = 1, 2, \ldots, n, \quad (4.1)$$

em que $\{q = 1, 2, \ldots, n_k\}$ identifica cada medida de similaridade $\hat{\gamma}_q(h_k)$ calculada em um particular estrato. Essa notação segue o mesmo critério em todos os estratos, pois apenas k determina o *lag* ou estrato e, além disso, note que, no geral, o índice $\{q = 1, \ldots, n_k\}$ não é uma forma de ordenar $\hat{\gamma}(h_k)$, mas, sim, de identificação. Porém, quando se tratar das estatísticas bootstrap (qualquer função das reamostras), o indexador q automaticamente ranqueia essas grandezas.

Com a Expressão (4.1) é possível calcular todas as combinações dos pares de semidiferenças quadradas

$$\begin{aligned}\boldsymbol{\Gamma}(h_k) = \{&lag_1 = [\hat{\gamma}_1(h_1), \hat{\gamma}_2(h_1), \ldots, \hat{\gamma}_{n1}(h_1)],\\ &lag_2 = [\hat{\gamma}_1(h_2), \hat{\gamma}_2(h_2), \ldots, \hat{\gamma}_{n2}(h_2)],\\ &lag_3 = [\hat{\gamma}_1(h_3), \hat{\gamma}_2(h_3), \ldots, \hat{\gamma}_{n3}(h_3)],\\ &\ldots \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots,\\ &lag_k = [\hat{\gamma}_1(h_k), \hat{\gamma}_2(h_k), \ldots, \hat{\gamma}_{n_k}(h_k)]\},\end{aligned} \tag{4.2}$$

que constituem o semivariograma experimental de nuvens, agrupadas devidamente no k-ésimo *lag* (ou estrato). Assim, o conjunto $\boldsymbol{\Gamma}(h_k)$ forma a base de dados fundamental para o desenvolvimento da medologia proposta neste trabalho científico, de modo geral.

Agora, simultaneamente, tornam-se as amostras aleatóarias bootstrap e aplica-se a função $\{\hat{\wp}_b(\Gamma^*(h_k)) : b = 1, \ldots, B\}$, gerando as estatísticas apropriadas. Estas são ordenadas crescentemente para gerar uma distribuição empírica da função $\hat{\wp}_b(h_k)$ aplicada nas B repetições bootstrap. Dessa forma, obtêm-se quaisquer quantis de interesse para se avaliar as incertezas do semivariograma *lag* a *lag*. Em outras palavras, realizou-se o bootstrap reamostrando cada quantidade disponibilizada por (4.2), dentro de cada particalar *lag*-k. Esse procedimento gera o conjunto de estatísticas baseadas na reamostragem,

$$\begin{aligned}\boldsymbol{\Gamma}^*(h_k) = \{&lag_1 = [\hat{\wp}_1(h_1), \hat{\wp}_2(h_1), \ldots, \hat{\wp}_B(h_1)],\\ &lag_2 = [\hat{\wp}_1(h_2), \hat{\wp}_2(h_2), \ldots, \hat{\wp}_B(h_2)],\\ &lag_3 = [\hat{\wp}_1(h_3), \hat{\wp}_2(h_3), \ldots, \hat{\wp}_B(h_3)],\\ &\ldots \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots,\\ &lag_k = [\hat{\wp}_1(h_k), \hat{\wp}_2(h_k), \ldots, \hat{\wp}_B(h_k)]\},\end{aligned} \tag{4.3}$$

determinadas pela quantidade B de repetições razoavelmente prefixada. A definição de B é relativa e varia de acordo com a estatística a ser estimada. Por exemplo, para se estimar a desvio padrão da média amostral de um processo por meio do bootstrap, Efron e Tibshirani (1993) sugerem ($25 \leq B \leq 200$). Já, quando se trabalha com correlação e teste de hipóteses, o número B de simulações seguro deve ser de, pelo menos, 1.000 repetições. Nesse estudo, testou-se o valor de B em larga escala, variando-o de 100 a 2.000. Ao conjunto de estatísticas estabelecidas em (4.3) denomina-se semivariograma de nuvens bootstrap. Finalmente, essa ideia pode ser sintetizada no seguinte algoritmo:

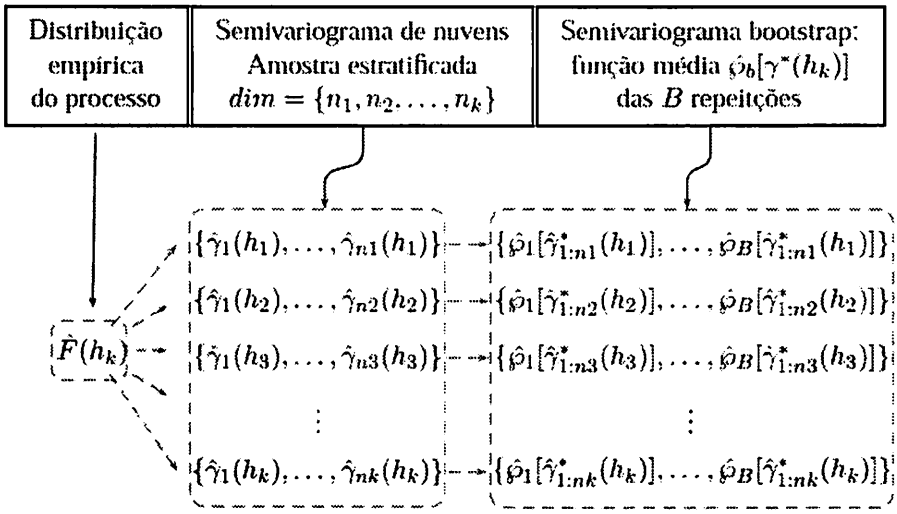


Figura 4.1 Algoritmo bootstrap para estimar o semivariograma de quantis, $\wp_b(h_k)$, de uma estatística $\hat{\gamma}(h_k)$. Teoricamente, cada amostra bootstrap é uma amostra aleatória independente de tamanho n_k gerada a partir de $\hat{F}(h_k)$

4.1.2 Usando o bootstrap na modelagem do semivariograma

Com o objetivo de avaliar as incertezas associadas ao semivariograma, apresenta-se o semivariograma de quantis bootstrap, conforme mostrado na Figura

4.2. A seguir, descreve-se a teoria para a contrução do semivariograma/variograma de quantis bootstrap. A metodologia será descrita em quatro passos, explorando tanto os aspectos teóricos quanto computacionais.

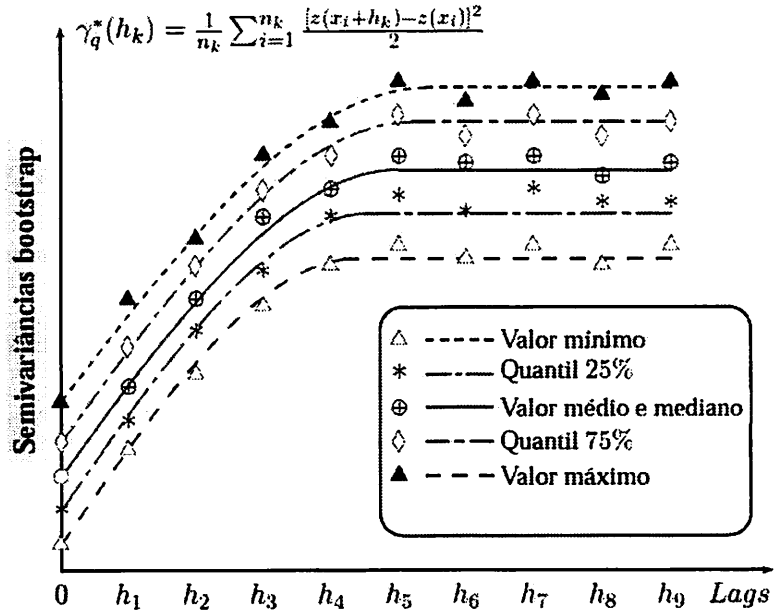


Figura 4.2 Semivariograma de quantis bootstrap $\{\gamma_q^*(h_k) : q = 1, \ldots, n_k\}$: os semivariogramas experimentais (representados pelos símbolos) foram ajustados a sentimento (linhas), a partir das B repetições bootstrap, considerando os quantis {mínimo, 25%, médio/mediano, 75% e máximo}, para cada *lag*. O intercepto *nugget* foi estimado naturalmente pelo modelo esférico, com base nos dados dos *lags* reamostrados

1º passo: Conforme mencionado, as medidas de continuidade espacial (semivariograma de nuvens) $\{\hat{\gamma}(h_k) = \frac{1}{2}[z(x_i + h_k) - z(x_i)]^2 : i = 1, \ldots, n; k = 1, \ldots, m\}$ são uma estatística que consiste de todos os possíveis pares de pontos, x_i, que estão localizados à mesma distânica h_k (isto é, aqueles pontos espaciais que estão a um mesmo *lag* distante ou, então, a uma mesma classe de distância). O índice k define o número de *lags* a serem consi-

derados na estimação do semivariograma de nuvens. O conjunto de dados $\Gamma(h_k)$, que captura o tipo de dependência espacial do processo estocástico $Z(x)$, será a base de dados para se aplicar a metodologia e construir o semivariograma de quantis bootstrap. Note que cada componente de $\Gamma(h_k)$ em (4.2) é vetor de dados com dimensões diferentes porque dependem da distância, h_k. Quanto maior for o *lag*, menor é o número de pares e vice-versa, alterando, assim, a dimensão dos vetores de $\Gamma(h_k)$. Esse fato dificulta o arranjo dos dados na forma matricial e, por isso, optou-se pelo apropriado formato lista (`list(objetos)`) do programa *R* para se efetuar a técnica de reamostragem computacionalmente. Perceba que as semivariâncias pertencentes a um específico *lag*, $\hat{\gamma}(h_K)$, forma a k-ésima partição amostral de $\Gamma(h_k)$. Em outras palavras, cada *lag* é considerado um estrato, como ocorre no sistema de amostragem estratificada. Parece bastante coerente a ideia de se tratar cada *lag* como sendo um estrato definido em função da distância (h_k), porque isso já ocorre naturalmente na estimação do semivariograma e, portanto, deseja-se apenas formalizar isso teoricamente para facilitar o entendimento da abordagem bootstrap aplicada nas medidas de similaridade espaciais, $\hat{\gamma}(h_k) = \frac{[z(x_i+h_k)-z(x_i)]^2}{2}$.

2º passo: Aplica-se uma quantidade B de repetições bootstrap, com reposição, em cada objeto — vetor amostral de semivariâncias — do conjunto $\Gamma(h_k)$, independentemente, gerando as amostras bootstrap $\hat{\gamma}_q^*(h_k)$ e, por sua vez, suas médias $(\hat{\wp}_b[\hat{\gamma}_q^*(h_k)])$ são calculadas. Assim, o objeto $\Gamma^*(h_k) = [\hat{\wp}_b(h_k)]$ é uma matrix de ordem $(B \times k)$, sendo o valor de B arbitrário e k dependente do número de lags disponíveis no semivariograma. Suponha, por exemplo, um processo imagiário que propiciou a estimação de um semivariograma de nuvens com 10 *lags*, $\Gamma(h_k) = \{\gamma_{1:n1}(h_1), \ldots, \gamma_{1:n10}(h_{10})\}$. Após 50

replicações bootstrap de $\Gamma(h_k)$, obtém-se o *array* de semivariâncias,

$$\Gamma^*(h_k) = \begin{vmatrix} \hat{\wp}_1[\hat{\gamma}^*_{1:n1}(h_1)] & \hat{\wp}_1[\hat{\gamma}^*_{1:n2}(h_2)] & \cdots & \hat{\wp}_1[\hat{\gamma}^*_{1:n10}(h_{10})] \\ \hat{\wp}_2[\hat{\gamma}^*_{1:n1}(h_1)] & \hat{\wp}_2[\hat{\gamma}^*_{1:n2}(h_2)] & \cdots & \hat{\wp}_2[\hat{\gamma}^*_{1:n10}(h_{10})] \\ \hat{\wp}_3[\hat{\gamma}^*_{1:n1}(h_1)] & \hat{\wp}_3[\hat{\gamma}^*_{1:n2}(h_2)] & \cdots & \hat{\wp}_3[\hat{\gamma}^*_{1:n10}(h_{10})] \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\wp}_{50}[\hat{\gamma}^*_{1:n1}(h_1)] & \hat{\wp}_{50}[\hat{\gamma}^*_{1:n2}(h_2)] & \cdots & \hat{\wp}_{50}[\hat{\gamma}^*_{1:n10}(h_{10})] \end{vmatrix}_{50\times 10} , \tag{4.4}$$

em que $\{k = 1, \ldots, 10\}$. De fato, cada linha de $\Gamma^*(h_k)$ corresponde a um semivariograma bootstrap $\hat{\wp}_b[\gamma^*_{1:nk}(h_k)]$ com 10 *lags* e cada coluna representa um estrato, h_k, que contém a distribuição das semivariâncias bootstrapped segundo o valor B.

3 º passo: O conjunto de "pseudodados" bootstrap, $\Gamma^*(\cdot)$ forma a base para os procedimentos de modelagem do semivariograma de quantis por meio dos métodos dos quadrados mínimos, em geral. Note que o nome pseudodados é utilizado para diferenciar dos dados originais, que são as variáveis regionalizadas do processo estocástico, $\{z(x) = z(x_1), z(x_2), \ldots, z(x_n)\}$. Assim, as semivariâncias são as nuvens de pontos a serem usadas para o ajuste dos modelos de suavização usuais em Geoestatística. Um exemplo típico pode ser verificado na Figura (4.2). Uma vez conhecida a distribuição empírica das semivariâncias, como definido em (4.4) do 2º passo, escolhem-se os quantis desejados pelos quais se processa o ajuste dos modelos mediante os Quadrados Mínimos Ponderados (QMP). Considerando a filosofia central da Geoestatística, a continuidade espacial de variáveis regionalizadas é uma função da distância e associando esse fato à simplicidade de implementação

computacional em relação aos outros sistemas de pesos possíveis, sugere-se o uso do critério de ponderação $\{W_{N/h} = \frac{N(h)}{h^p} : p = 1, 2, \ldots\}$. Diferentes métodos de ajuste de semivariograma estão descritos em Diggle e Ribeiro Junior (2007); Schabenberger e Gotway (2005); Cressie (1993). O semivariograma, em geral, foi modelado otimizando-se o conjunto de parâmetros de covariância, $\boldsymbol{\theta}$, que minimiza a função objetivo

$$\sum_{h=1}^{K} W_{N/h}[\varepsilon_h]^2 = \sum_{h=1}^{K} W_{N/h}[\hat{\gamma}(h) - \gamma(h,\boldsymbol{\theta})]^2 \tag{4.5}$$

em que $\hat{\gamma}(h)$ é o semivariograma experimental (pseudodados: pode ser pensado como uma variável resposta, comum em regressão) e $\gamma(h,\boldsymbol{\theta})$ denota o modelo teórico em função do vetor $\boldsymbol{\theta} = (\tau^2, \varrho^2, \varphi)$.

Segue um pequeno exemplo demonstrativo do procedimento bootstrap aplicado no semivariograma *cloud*, conforme a metodologia descrita acima.

Exemplo didático 2

Dados espaciais: gera-se um campo aleatório gaussiano por meio da função `grf` da *geoR* (RIBEIRO JUNIOR; DIGGLE, 2001)

```
set.seed(321312)
d.GaussRF <- grf(9,grid="reg",cov.pars=c(0.4,0.25),
+ nug=0.1,mean=10,cov.model="sph",RF=TRUE)
print(d.GaussRF) # não apresentados
```

Pseudodados: calculando o semivariograma *cloud*

```
(Gama.h <- variog(d.GaussRF,option="bin",max.dist=1,
+ bin.cloud=TRUE)
Gamma.cloud <- list(Lag1=list.cloud[[1]],
+ Lag2=list.cloud[[2]],Lag3=list.cloud[[3]],
+ Lag4=list.cloud[[4]],Lag5=list.cloud[[5]])
print(Gama.h) # visualisando os objetos (saída da geoR)
$ Lag1
[1] 0.66194 0.25092 0.10741 0.13962 0.62617 0.88522
0.43859 0.55562 0.01079 0.08005 0.14628 0.32074
$ Lag2
[1]  0.193550 1.727946 1.252280 0.002108 1.091459
0.077615 0.213878 0.721258
$ Lag3
[1]  0.23606 0.02603 0.07278 0.25845 0.03820 0.03381
$ Lag4
[1] 0.093303 0.295733 0.425429 0.699102 0.973718
0.003359 0.228857 0.217906
$ Lag5
[1] 0.0005074 0.1053052
```

Construção de $\Gamma(h)$: capturando os pseudodados (semivariograma *cloud*) do objeto Gama . h e, organizando-os em estratos amostrais, $Lag_1, \ldots, Lag_5$, formando a base de dados armazenada na lista $\Gamma(h)$, ultilizada para gerar as reamostras segundo a metodologia bootstrap.

$$\Gamma(h) = \{lag_1 = [0{,}66194; 0{,}25092; 0{,}10741; 0{,}13962; 0{,}62617; 0{,}88522,$$
$$0{,}43859; 0{,}55562; 0{,}01079; 0{,}08005; 0{,}14628; 0{,}32074]_{n_1=12},$$
$$lag_2 = [0{,}193550; 1{,}727946; 1{,}252280; 0{,}002108; 1{,}091459;$$
$$0{,}077615; 0{,}213878; 0{,}721258]_{n_2=8},$$
$$lag_3 = [0{,}23606; 0{,}02603; 0{,}07278; 0{,}25845; 0{,}03820; 0{,}03381]_{n_3=6},$$
$$lag_4 = [0{,}093303; 0{,}295733; 0{,}425429; 0{,}699102; 0{,}973718; 0{,}003359;$$
$$0{,}228857; 0{,}217906]_{n_4=8},$$
$$lag_5 = [0{,}0005074; 0{,}1053052]_{n_5=2}\}$$

Reamostragem bootstrap: repetições bootstrap $\{\mathbf{R_b}, b = 1, \ldots, 8\}$ obtidas particularmente em cada um dos 5 estratos (*lags*) de dimensões $\{n_1 = 12, n_2 = 8, n_3 = 6, n_4 = 8, n_5 = 2\}$, a partir da lista de objetos $\Gamma(h_k)$ (semivariâncias ou dissimilaridades). Perceba que o método abordado considera as diferentes dimensões de cada estrato. O código R para realizar as repetições bootstrap (com reposição) está disponível abaixo, seguido dos resultados obtidos e exibidos na forma matricial. Ao avaliar essas matrizes, nota-se, facilmente, que alguns valores aparecem várias vezes na reamostra, enquanto outros não foram contemplados pelo sorteio aleatório.

```
set.seed(2321321)
n.boots <- 8 # número de repeticões
n.bins <- length(v.cloud$u) # número de lags
mu.star <- matrix(0,nrow=n.boots,ncol=n.bins)
me.star <- matrix(0,nrow=n.boots,ncol=n.bins)
index.boot <- vector(mode="numeric")
```

```
list.cloud <- list  (mode="numeric")
Lag1 <- list(mode="numeric") # estrato 1
Lag2 <- list(mode="numeric") # estrato 2
Lag3 <- list(mode="numeric") # estrato 3
Lag4 <- list(mode="numeric") # estrato 4
Lag5 <- list(mode="numeric") # estrato 5
for(i in 1:n.boots){
for(j in 1:n.bins){
index.boot <- sample(1:length(v.cloud$bin.cloud[[j]]),
+ length(v.cloud$bin.cloud[[j]]),replace=T)
if (j==1){
R.Lag1[[i]] <- v.cloud$bin.cloud[[j]][index.boot]}
if (j==2){
R.Lag2[[i]]  <- v.cloud$bin.cloud[[j]][index.boot]}
if (j==3){
L3[[i]]  <- v.cloud$bin.cloud[[j]][index.boot]}
if (j==4){
R.Lag4[[i]] <- v.cloud$bin.cloud[[j]][index.boot]}
if (j==5){
R.Lag5[[i]] <- v.cloud$bin.cloud[[j]][index.boot]}
mu.star[i,j] <-
+ mean(v.cloud$bin.cloud[[j]][index.boot])
me.star[i,j] <-
+ median(v.cloud$bin.cloud[[j]][index.boot])
list.cloud[[j]] <- v.cloud$bin.cloud[[j]]
}}
```

$Lag_1 =$

n_1	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
1	0,439	0,107	0,107	0,107	0,146	0,626	0,321	0,080
2	0,626	0,556	0,439	0,662	0,140	0,140	0,885	0,439
3	0,885	0,107	0,140	0,662	0,885	0,321	0,626	0,321
4	0,011	0,885	0,662	0,556	0,626	0,080	0,439	0,885
5	0,011	0,321	0,556	0,080	0,626	0,439	0,251	0,885
6	0,626	0,321	0,080	0,251	0,439	0,439	0,885	0,556
7	0,146	0,321	0,107	0,251	0,662	0,321	0,080	0,080
8	0,885	0,080	0,011	0,107	0,080	0,080	0,626	0,140
9	0,080	0,251	0,556	0,080	0,080	0,011	0,885	0,080
10	0,146	0,321	0,080	0,662	0,251	0,146	0,251	0,146
11	0,662	0,011	0,321	0,321	0,107	0,140	0,885	0,011
12	0,321	0,321	0,885	0,439	0,662	0,885	0,321	0,556

$Lag_2 =$

n_2	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
1	0,214	1,728	1,091	1,091	1,091	1,252	0,721	0,194
2	0,721	0,078	1,091	0,214	1,091	1,728	0,002	0,002
3	0,214	0,214	0,078	1,252	0,078	0,002	0,078	1,728
4	1,252	0,721	0,721	0,002	0,078	0,078	0,078	0,078
5	1,091	0,194	0,721	1,728	0,721	1,728	1,728	0,214
6	0,078	0,721	0,002	1,091	0,194	1,091	0,214	0,078
7	0,214	0,078	0,721	0,078	1,252	1,252	1,091	1,728
8	1,091	0,214	0,721	0,002	1,091	0,214	1,252	0,721

$Lag_3 =$

n_3	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
1	0,073	0,073	0,258	0,034	0,034	0,236	0,034	0,258
2	0,073	0,038	0,236	0,034	0,258	0,258	0,038	0,034
3	0,073	0,258	0,258	0,034	0,026	0,073	0,038	0,073
4	0,258	0,073	0,073	0,258	0,236	0,038	0,034	0,034
5	0,034	0,026	0,026	0,258	0,236	0,073	0,026	0,034
6	0.073	0.026	0,038	0,236	0,026	0,038	0,258	0,073

$Lag_4 =$

n_4	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
1	0,218	0,218	0,003	0,296	0,699	0,699	0,699	0,003
2	0,003	0,003	0,093	0,974	0,218	0,229	0,218	0,229
3	0,974	0,425	0,218	0,425	0,699	0,218	0,296	0,974
4	0,229	0,974	0,229	0,229	0,229	0,296	0,229	0,296
5	0,699	0,218	0,974	0,296	0,229	0,425	0,093	0,003
6	0,296	0,699	0,229	0,218	0,229	0,425	0,974	0,699
7	0,093	0,003	0,229	0,699	0,699	0,296	0,003	0,699
8	0,003	0,093	0,296	0,229	0,218	0,093	0,296	0,229

$Lag_5 =$

n_5	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
1	0,105	0,105	0,001	0,001	0,001	0,105	0,001	0,105
2	0,105	0,001	0,105	0,001	0,105	0,105	0,105	0,001

Média das repetições bootstrap: agora, calcula-se a média das repetições estratificadas, exibidas no item anterior, obtendo-se, então, as semivariâncias bootstrap ($\hat{\gamma}^*(h)$). Esses resultados estão alocadas no *array* $|\hat{\Gamma}^*(h)|_{8\times5}$, abaixo. Note que foram realizadas 8 repetições em 5 estratos, gerando uma matriz de dimensão 40. Observe que, por facilidade, a média e a mediana já

foram calculadas, automaticamente, pelo algoritmo acima e foram armazenadas no objetos `mu.star` e `me.star`, respectivamente.

$$\boldsymbol{\Gamma}^*(h) = \begin{vmatrix} B_b & \mathbf{Lag_1} & \mathbf{Lag_2} & \mathbf{Lag_3} & \mathbf{Lag_4} & \mathbf{Lag_5} \\ \mathbf{R_1} & 0{,}403 & 0{,}609 & 0{,}097 & 0{,}314 & 0{,}105 \\ \mathbf{R_2} & 0{,}300 & 0{,}493 & 0{,}082 & 0{,}329 & 0{,}053 \\ \mathbf{R_3} & 0{,}329 & 0{,}643 & 0{,}148 & 0{,}284 & 0{,}053 \\ \mathbf{R_4} & 0{,}348 & 0{,}682 & 0{,}142 & 0{,}421 & 0{,}001 \\ \mathbf{R_5} & 0{,}392 & 0{,}700 & 0{,}136 & 0{,}402 & 0{,}053 \\ \mathbf{R_6} & 0{,}302 & 0{,}918 & 0{,}119 & 0{,}335 & 0{,}105 \\ \mathbf{R_7} & 0{,}538 & 0{,}646 & 0{,}071 & 0{,}351 & 0{,}053 \\ \mathbf{R_8} & 0{,}348 & 0{,}593 & 0{,}084 & 0{,}392 & 0{,}053 \end{vmatrix}_{8\times 5} . \quad (4.6)$$

A matrix $\boldsymbol{\Gamma}^*(h)$ em (4.6) contém a distribuição empírica das médias *bootstrapped* ou, equivalentemente, tem-se a dispersão das semivariâncias estratificadas em suas respectivas classes de distâncias $\{Lag_1, Lag_2, Lag_3, Lag_4, Lag_5\}$. Essa estrutura permite a construção do semivariograma de quantis bootstrap, conforme apresentado a seguir.

Quantis bootstrap: a partir dos resultados do item anterior ($\boldsymbol{\Gamma}^*(h)$) reamostrados de $\boldsymbol{\Gamma}(h)$, podem-se obter quaisquer quantis das semivariâncias bootstrap e construir o semivariograma percentil bootstrap. Na Tabela 4.1 são mostrados os resultados desse procedimento, obtidos pelo simples comando:

```
q.boots <- sapply(mu.star,quantiles).
```

Na Tabela 4.2 está sintetizo o resultado final de todo o procedimento bootstrap necessário para a construção do semivariograma de quantis, porém, pode-se

Tabela 4.1 Percentis bootstrap originados da matrix 4.6.

	0%	25%	50%	75%	100%
Lag_1	0,3000933	0,3219927	0,3481407	0,3948196	0,5379305
Lag_2	0,4933746	0,6052821	0,6444894	0,6866642	0,9181889
Lag_3	0,0714189	0,0837764	0,1083233	0,1376545	0,1483302
Lag_4	0,2838238	0,3255495	0,3430753	0,3942486	0,4206670
Lag_5	0,0005074	0,0529063	0,0529063	0,0660061	0,1053052

obter quaisquer intervalos de confiança de interesse do pesquisador. Portanto, essa mesma estrutrura de dados é também o suporte para a formalização do teste de normalidade multivariada para Geoestatística, proposto neste trabalho científico. Por se tratar de um pequeno exemplo didático, para melhor compreender os procedimento bootstrap aplicado ao semivariograma de nuvens, não se apresentoram os gráficos ilustrativos nesse caso, tendo em vista o semviariograma experimental possuir muito pouco pontos. Mas, em geral, há vários exemplos gráficos possíveis de se obter por meio dessa metodologia, nas seções a seguir. A seguir descreve-se os códigos para a obtenção dos dados da Tabela 4.2.

```
perc2.5   <- vector (mode="numeric") # percentil 2,5%
perc97.5 <- vector (mode="numeric") # percentil 97,5%
for (k in 1:n.bins){
perc2.5[k]  <- quantile(gama.star[,k],0.025)
perc97.5[k] <- quantile(gama.star[,k],0.975)}
# imprime formato tabel (# requer o pacote xtable):
(gama.boots <- xtable((cbind(perc2.5,perc97.5))))
```

Tabela 4.2 Distribuição das semivariâncias bootstrap originadas da matrix 4.6, em função da classe de distância (h) e $N(h)$ é o número de pares por *lag*.

Lag	h	$N(h)$	$\hat{\gamma}_{min}$	$\hat{\gamma}_{2,5\%}$	$\hat{\gamma}_{50\%}$	$\bar{\gamma}$	$\hat{\gamma}_{97,5\%}$	$\hat{\gamma}_{max}$
Lag_1	0,49	12	0,30	0,30	0,30	0,37	0,51	0,54
Lag_2	0,71	08	0,49	0,51	0,56	0,66	0,88	0,92
Lag_3	1,03	06	0,07	0,07	0,07	0,11	0,15	0,15
Lag_4	1,14	08	0,28	0,29	0,25	0,35	0,42	0,42
Lag_5	1,36	02	0,00	0,01	0,05	0,06	0,11	0,11

Contudo, a forma tradicional de se representar a distribuição empírica bootstrap é por meio do histograma das estatísticas obtidas, isto é, as semivariâncias no caso do semivariograma. Assim, com fins demostrativos, apresenta-se (veja Figura 4.3) apenas o histograma das semivariâncias do primeiro estrato ou *lag* bootstrapped.

4.1.3 Teste gaussiano para Geoestatística (TGGeo)

Dada a importância de se identificar um processo gaussiano, apresenta-se um teste de normalidade multivariada para os dados de Geoestatística (isto é, dados contínuos e espacialmente correlacionados) ou, simplesmente, Teste Gaussiano para Geoestatística (TGGeo). Em suma, trata-se de uma ferramenta exploratória e investigativa que torna viável e mais atrativa a abordagem da Geoestatística Baseada em Modelo (DIGGLE; RIBEIRO JUNIOR, 2007; DIGGLE; TAWN; MOYEED, 1998), por ser capaz de avaliar se um dado processo estocástico possui distribuição normal multivariada.

A ideia central é usar o semivariograma como base para gerar uma es-

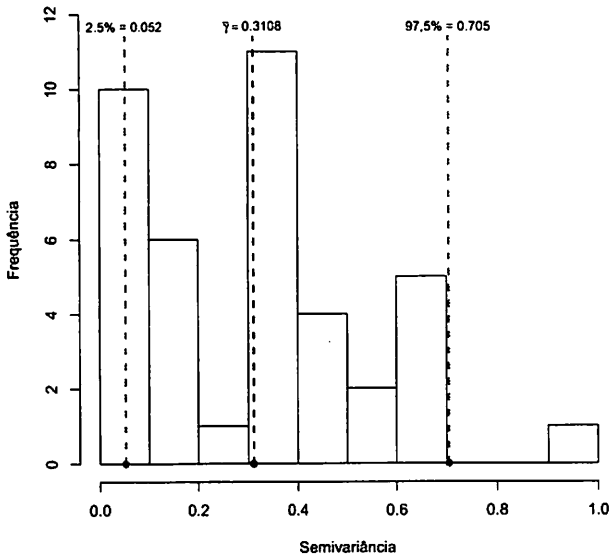


Figura 4.3 Histograma das repetições bootstrap do primeiro estrato ou *lag*

tatística de teste. O suporte teórico para isso está diretamente relacionado com os tradicionais métodos de ajustes do modelo de semivariograma, $\gamma(h,\theta)$, associado ao semivariograma experimental $\hat{\gamma}(h)$. Partindo dessa lógica, parece bastante razoável confrontar as propriedades inerentes aos métodos baseados na máxima verossimilhança (MV) e a dos métodos dos quadrados mínimos (QM) no intuito de avaliar como eles se comportam, conjuntamente, em presença de normalidade multivariada nos dados de um processo espacialmente contínuo. Sob normalidade, quadrados mínimos e máxima verossimilhança levam a estimativas semelhantes. Sabe-se que os ajustes com base nos métodos dos quadrados mínimos não requerem normalidade entre as amostras, mas, quando se utilizam os ajustes baseados em verossimilhança, essa normalidade (ou outro pressuposto distribucional qualquer) é requerida. Assim, considerando o princípio estatístico característico de

ambos os métodos, é possível criar uma estatística de teste capaz de quantificar o grau de semelhança entre aqueles ajustes, a saber valor de plausibilidade (valor-pl = δ) e, portanto, aceitar ou rejeitar a hipótese, previamente postulada, de os dados espaciais seguirem distribuicão normal multivariada, isto é, a hipótese H_0: os dados são gaussianos. A quantidade δ determina que porcentagem das estimativas obtidas pelo critério da máxima verossimilhança pertencem ao intervalo de confiança bootstrap e será adotado um valor de plausibilidade $\delta \leq 40\%$ como critério para se rejeitar H_0, isto é, os dados não são gaussianos e, obviamente, aceita-se H_0 com uma estatística $(1 - \delta > 60\%)$.

Construção do TGGeo

Basicamente, o TGGeo é realizado por meio de análise gráfica que verifica se ambos métodos de ajustes, QM e MV, têm o mesmo desempenho quando se estimam os parâmetros de covariância (pepita, contribuição e alcance) do semivariograma experimental. Em suma, o TGGeo testa, com algum critério, se as curvas de "regressão" pertencem ao intervalo de confiança percentil bootstrap construído com um nível α de significância. O teste pode ser formalizado da seguinte modo:

1. estabelecem-se as hipóteses H_0 : o campo aleatório é gaussiano e H_a : o campo aleatório tem outra distribuição diferente da gaussiana. Outra forma de construção também é válida:

$$\text{Hipóteses} = \begin{cases} H_0 : \text{QM} = \text{MV} \\ H_a : \text{QM} \neq \text{MV} \end{cases} ;$$

2. executa-se a reamostragem nos pseudodados $(\hat{\gamma}(h_k))$ do objeto $\boldsymbol{\Gamma}(h_k)$, seguindo as instruções dadas na Secão 4.1.1 (mais precisamente os passos 1

e 2), para a obtenção do intervalo de confiança bootstrap com $1 - \alpha$ de confiança. Em todas as aplicações desta tese foi fixado um valor nominal $\alpha = 0{,}05$ para realizar o teste TGGeo. Porém, pode-se adotar qualquer valor de interesse para α. Note que, nesse estágio, além do semivariograma padrão estimado pelo bootstrap (idêntico ao proposto por Matheron), obtiveram-se as semivariâncias relativas ao percentil 2,5% ($q_{0,25}$) e 97,5% ($q_{0,975}$), gerando as bandas inferior e superior, respectivamente, do intervalo de confiança percentil bootstrap (RIZZO, 2008; CHERNICK, 2008);

3. consumado o item anterior, suavizaram-se as bandas do intervalo de confiança, $[q_{\frac{\alpha}{2}}, q_{(1-\frac{\alpha}{2})}]$, usando a ferramenta de modelagem de curvas não parametricas sobre um *scatterplot*, implementada na função `loess` do programa *R* — versão contemporânea da função `lowess`, definida como um filtro de curva não-paramétrico (CRAWLEY, 2006, p. 151). Esse ajuste é de suma importância porque dá um caracter contínuo às discretas semivariâncias que formam as bandas de confiança estimada pelo bootstrap. Com isso, cria-se um método para avaliar que proporção da curva de MV figurou dentro dos extremos de confiança, respeitando o valor nominal α prefixado;

4. estimadas as banda de confiança, modela-se o semivariograma padrão pelos métodos da MV e dos QM, o que é uma prática comum para os analistas em Geoestatística e bastante explorada nas literaturas. Certamente, atendendo às pressuposições básicas da Geoestatística e ajustando-se o melhor modelo de ambos os métodos mencionados, o pesquisador tem em mãos elementos plausíveis que o habilitam a decidir a favor ou contra a hipótese de H_0, conforme postulado no item 1. A decisão é tomada com base na estatística de teste (valor-pl) que é estabelecida por meio da proporção de valores preditos que estiverem no dentro das bandas de confiança, fixadas segundo um

valor crítico α. Mas, deve-se esclarecer que, por questões óbivas, o TGGeo proposto avalia as curvas somente até ao *alcance prático* do modelo. Esse fato é enfatizado pelas linhas contínuas da bandas de confiança e, caso contrário, as linhas são pontilhadas (Figuras 4.4 e 4.5). Porém, em caso cruciais, é facultado ao usuário relaxar o critério adotado e tomar a decisão subjetivamente. Esse afrouxamento do critério é permitido porque, em determinados fenômenos, ou o padrão dependência espacial não é tão claro, ou ocorre de as semivariâncias não se estabilizarem conforme é esperado, o que afeta o comportamento das curvas, principalmente no que diz respeito ao método da MV. Vale lembar que a Geoestatística não é apenas um ajuste de pontos em um *scatterplot*, como em regressão. Pode-se dizer que se faz modelagem em Geoestatística com bastante interatividade e a subjetividade é importante, embora tenham-se, atualmente, recursos computacionais bastante sofisticados para tal. Nesses casos, a experiência e o bom censo do pesquisador são preponderantes (é mais influente) na sua decisão. É bom que o leitor compreenda que, em qualquer situação, o valor plausível sempre funciona como um termômetro, eficiente e suficiente, sinalizando a evidência dos fatos. Mas, nesses casos críticos, dúbios, o usuário respaldado pelas evidências apontadas pelo TGGeo, juntamente com sua prática, tem poder de decisão, livremente.

As Figuras 4.4 e 4.5 ilustram, distintamente, um esquema geral do TGGeo, ratificando as únicas proposições dicotômicas cabíveis para o teste, a tomar conhecimento: 1) os dados seguem uma distribuição normal multivariada e 2) os dados têm outra distribuição diferente da gaussiana. Analisando-se a Figura 4.4, pode-se verificar que o momento de inércia do critério de ajuste MV em relação ao critério QM não foi significativo e, nesse caso, o TGGeo acusa normalidade

multivariada nos dados. Colocando de outra forma: o modelo baseado em MV pertence ao intervalo de confiança bootstrap com um nível crítico aceitável, por exemplo, com um risco valor-pl=$\delta \times 100\%$. Contrariamente, a Figura 4.5 representa uma situação em que o TGGeo fornece evidências de não normalidade multivariada na amostra. Visivelmente, mais de $\delta = 40\%$ do modelo baseado em MV está fora do (*off-side*) dos limites de segurança, isto é, eles se comportam de forma diferente para o mesmo processo. Resumidamente, o princípio básico desse teste consiste em informar se o ajuste feito sobre os dados reais (MV) e o ajuste feito sobre a nuvens de pontos do semivariograma (método dos QM) predizem o padrão espacial do mesmo campo aleatório (população), isto é, o mesmo semivariograma. De fato, deve-se avaliar a inércia (deslocamento) do critério da máxima verossimilhança a partir do ajuste dos QM. Isso se deve ao fato de o método dos QM estimar os parâmetros de um modelo descrevendo a média de um vetor aleatório (SCHABENBEGER; GOTWAY, 2005), fazendo com que a linha preditiva se ajuste perfeitamente aos pontos do gráfico, o que difere bastante do dos procedimentos da MV. Portanto, esse comportamento permite considerá-la em repouso absoluto, tendo em vista ser o modelo realmente de quadrados mínimo, é claro.

Note que o modelo ajustado pelo método da máxima verossimilhança deve ser avaliado somente até ao alcance real (delimitado pelo linha contínua vermelha), enquanto o ajuste por quadrados mínimos (denominado de modelo completo) é considerado em todo o semivariograma (Figuras 4.4 e 4.5). Além disso, convém esclarecer que o símbolo $\oplus$ apresentado em todos os gráficos do TGGeo é formado pela combinação de dois outros símbolos e não constitui um única simbologia: o símbolo ($\circ$) representa o valor médio de cada estrato e o símbolo (+) expressa a médiana desses mesmos estratos. Essa simbologia pode ser usada porque a média e a mediana são praticamente iguais dentro de casa classe de distância avaliada

como estatísticas geradas pelo bootstrap.

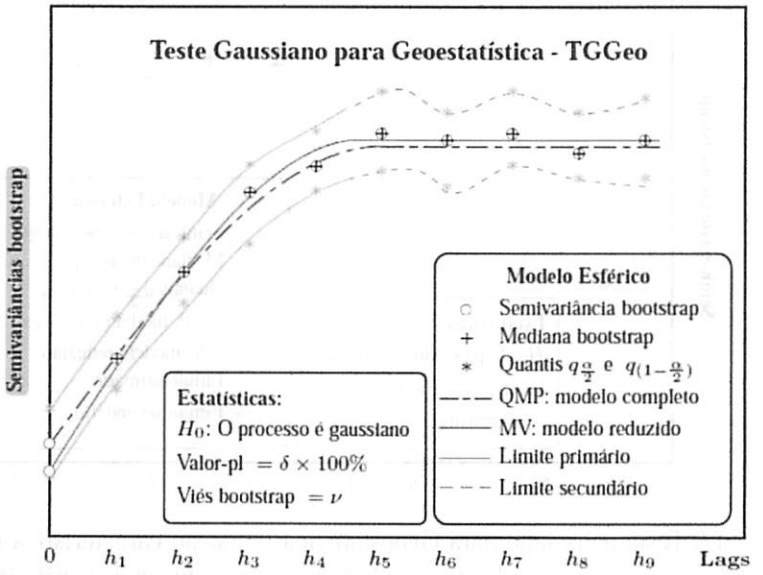


Figura 4.4 Teste gaussiano para Geoestatística: a gravura apresenta um comportamento padrão do TGGeo quando há normalidade, indicando um forte indício de os dados seguirem uma distribuição gaussiana

4.1.4 Cálculo do valor de plausibilidade

Gerou-se um vetor $\mathbf{h}'$ com 2.000 valores preditores h, sendo zero o seu menor valor e fechando com a distância máxima, $\mathbf{h}' = (0, \ldots, h_{max})'$, do semivariograma padrão estimado. Esses valores são úteis para traçar as curvas (CRAWLEY, 2006) do modelo de predição (linhas suavizadas) e dos limites inferior e superior do intervalo de confiança percentil. Com esse procedimento, tem-se um efeito (falsa impressão) de continuidade, mas, na verdade, trata-se de uma linha segmentada, formada por um conjunto finito, $\{h_1, h_2, \ldots, h_{2000}\}$, altamente denso de pixels, os quais denominam-se unidades preditoras. Essa segmentação

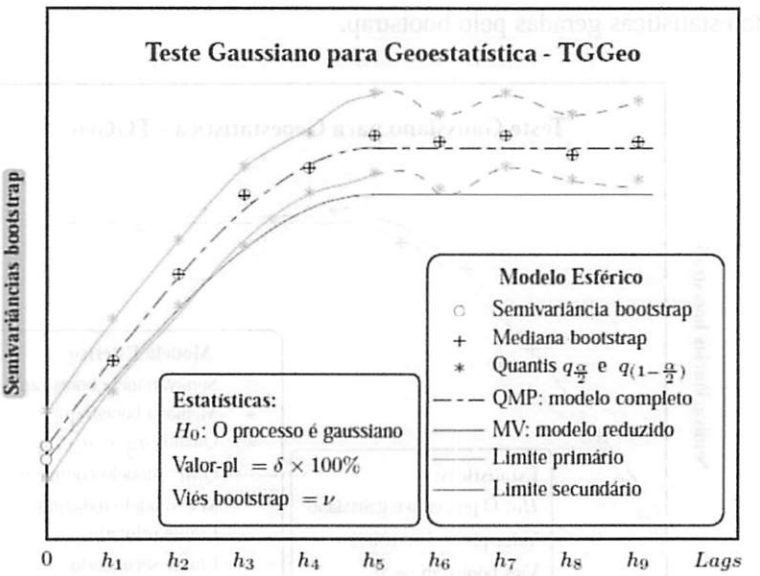


Figura 4.5 Teste gaussiano para Geoestatística: situação característica de rejeição de H_0, sugerindo que os dados não seguem uma distribuição normal multivariada

está representada no Quadro 4.6, pela projeção ampliada de parte dos pixels de um modelo de predição.

Por meio dessas unidades preditoras, o modelo teórico calcula 2.000 valores que compõem a linha predição. Da mesma forma, a função `loess` estima os 2.000 valores que constroem as bandas de confiança suavizadas, como pode ser visto nos Quadros 4.4 e 4.5. A partir dessas estatísticas, simplesmente, o valor-pl é encontrado contando a proporção de valores estimados pelo modelo teórico que figurarem dentro das bandas de confiança, conforme já mencionado.

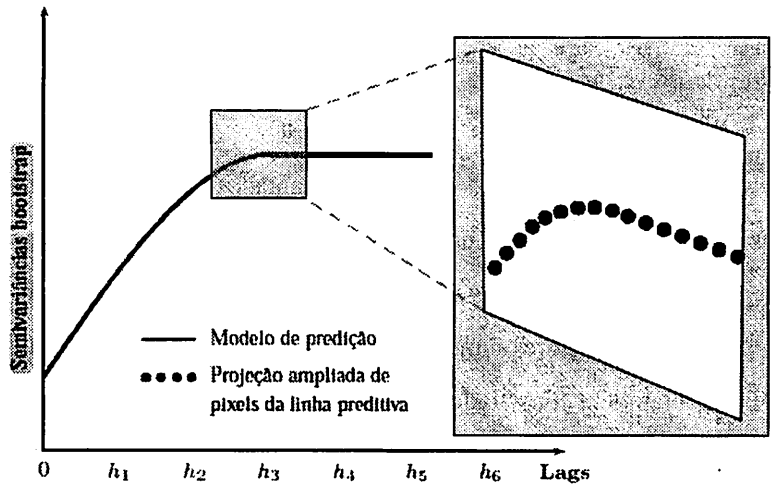


Figura 4.6 Construção do valor crítico do teste gaussiano para Geoestatística: esquema mostrando que linha preditiva é formada por adensameto de valores preditos (pixels). A projeção apresenta uma parte do modelo esférico simulado, ampliada, ratificando o fato de a linha preditiva ser discretizada em pixels, que são as unidades da linha preditiva

4.1.5 Critério de ponderação para ajustar o semivariograma por quadrados mínimos

Sugere-se um novo critério de pesos para suavizar o semivariograma experimental estimado, pelo critério dos quadrados mínimos e o seu comportamento pode ser comparado com o usual critério que utiliza o número de pares de pontos ($N(h)$), conforme pode ser visto na Figura 4.7. Antes de tudo, é importante esclarecer que os pesos exibidos no gráfico da Figura 4.7 foram reescalados para serem acomodados juntamente com a escala das semivariâncias. O objetivo é de apenas mostrar o comportamente dos respectivos pesos, visualizando-os de acordo com os *lags*. O sistema de ponderação é definido como

$$W_{N/h} = \frac{N(h_k)}{(h_k)^d},$$

em que $(h_k)^d$, com $\{d = 1, 2, \ldots\}$, é o valor da potência da distância no k-ésimo *lag* e $N(h_k)$ descreve o número de pares de distâncias obtidos nesses respectivos *lags*.

Esse gráfico comparativo dos pesos apresentados (Figura 4.7) é relativo à variável cobre, pertencente ao conjunto de dados do rio *Meuse*, coletados sobre um *lattice* irregular. Em dados com esse tipo de *lattice*, não se pode garantir que os primeiros *lags* tenham maiores números de pares, conforme destacado na Figura 4.7. Esse fato foi também constatado em outras análises, inclusive em dados sobre *lattice* regular, em que se fez variografia direcional. Perceba que, em geral, quando se fazem os incrementos (defasagens em função da distância h), espera-se obter o maior número de pares no primeiro *lag* e, depois, um decréscimo desse número na medida que a defasagem aumenta; mas isso não ocorreu para esse particular resultado apresentado.

Assim, pensando na qualidade e na simplicidade do ajuste do semivariograma por meio do critério dos quadrados mínimos ponderados (QMP) é que se estudou essa nova forma de ponderação, $W_{N/h}$. Isso porque dois problemas evidentes surgem ainda na fase da caracterização do padrão espacial subjacente ou seja, na estimação do semivarigrama experimental: (i) o problema da heterocedasticidade da variância devido ao número de pares $N(h)$ serem diferentes em cada classe de distância e esse fato torna o método dos quadrados mínimos ordinários (QMO) inadequado para o ajuste de semivariograma. Cressie (1985); Schabenberger e Gotway (2005) apresentaram uma boa discussão sobre o assunto; (ii) a extremidade final do semivariograma (últimos *lags*) é ponto crítico no processo de modelagam variográfica, devido aos baixos números de pares (LITTEL et al., 1996). Assim, tem-se o propósito de utilizar o método dos QMP incorporando, simultaneamente, essas duas questões problemáticas no seu procedimento

de ajuste. Outro fato importante que está bastante relacionado com qualidade do ajute do semivariograma experimental diz respeito ao seguinte princípio da teoria das variáveis regionalizadas (WACKERNAGEL, 2003):

> "A amostra georreferenciada é ao todo correlacionada, porém, os indivíduos mais próximos são mais semelhantes e essa semelhança muda com a variação da distância".

O critério de pesos $W_{N/h}$ transfere essa propriedade fundamental da Geoestatística para o modelo ajustado. Por esse motivo, podem-se gerar pesos que têm uma tendência linearmente decrescente (aproximadamente) na parte inicial do semivariograma e, após atingir o alcance, esse decaimento torna-se suave, tendendo a estabilizar exponencialmente (Figura 4.7). Note ainda, nessa mesma Figura, o comportamento espelhado (traçados côncavo e convexo), invertido, do critério de peso proposto $W_{N/h}$ em relação às semivariâncias, indicando uma lógica perfeita dos modelos com patamar. Assim, em suma, o que se tem é um conjunto de pesos suavizados que, ao mesmo tempo, considera a heterocedasticidade intra *lags* e corrige as flutuações problemáticas bruscas das semivariâncias além do alcance, penalizando-as por atribuir baixos valores de pesos. Além disso, com esse critério, sempre o primeiro *lag* receberá maior peso, independentemente do *lattice* adotado; fato esse, que o torna adequadamente ótimo.

O critério de pesos sugerido não é o foco deste trabalho. Trata-se de uma pequena contribuição para a ciência que surgiu, naturalmente, quando se desenvolvia a metodologia central desse estudo, o Teste Gaussiano para Geoestatística (TGGeo). Portanto, não se tem, nesse momento, interesse em avaliar o desempenho do critério de pesos $W_{N/h}$, o que será feito posteriormente.

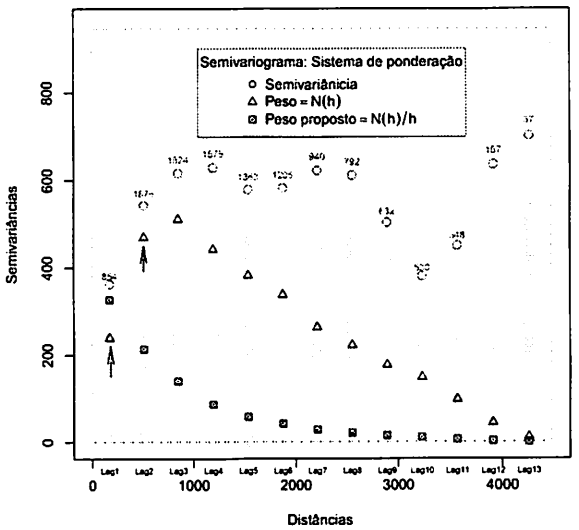


Figura 4.7 Demostrativo dos critério de peso: (△) comportamento dos pesos número de pares de diferenças quadráticas, $N(h)$ e (⊠) comportamento do sistema de ponderação proposto: razão entre $N(h)$ e a distânca h ($W_{N/h}$). Os valores exibidos no topo das hastes são os números de pares de pontos, enquanto o comprimento das hastas correspondem aos valores da semivariâncias

4.1.6 Semivariograma de quantis bootstrap

Na Figura 4.8 apresenta-se a forma característica do semivariograma de quantis proposto para avaliar os fenômemos de interesse com mais riqueza de detalhes, possibilitando avaliar as incertezas inerentes à estimação da estrutura de correlação espacial. A análise demonstrativa foi realizada nos dados do rio *Meuse* (*Utrecht*, 1993), para exemplificação — perceba que as simbologias dos valores mediano (+) e médio (o) do semivariograma de quantis estão sobrepostos, devido ao fato de seus valores serem muito próximos. O gráfico é uma ferramenta importante que somente foi possível mediante o uso das técnicas bootstrap, viabilizada

pela função *sample*, do programa R (RIZZO, 2008; CRAWLEY, 2006).

Sabe-se que o semivariograma experimental de nuvens é um poderoso estimador de processos espacialmente contínuos, que revela com integridade, confiando na amostragem, o real panorama do processo ocorrendo no campo aleatório, $\mathcal{R}$, de estudo (ISAAKS; SRIVASTAVA, 1989). Todavia, o mais comum na prática é usar o estimador clássico de Matheron para se detectar qual o tipo de padrão espacial que está ocorrendo e, com base neste, ajustar um modelo adequado para descrecer ou predizer o fenômeno. A justificativa dessa preferência está na facilidade em se trabalhar com a semivariância média de cada *lag* ou classe de distância, a qual é uma interessante estatística de resumo do padrão de dependência espacial, sem dúvida. Consequentemente, o conjunto de medidas de dissimilaridades (semivariograma dos pares das diferenças quadráticas) passa despercebido pelo analista em Geoestatística, infelizmente. Esse fato é reforçado (ou confirmado) principalmente pela maioria dos programas de computador que induzem o usuário a realizar somente a análise variográfica padrão, por ser a única opção disponível, lamentavelmente. Em vista disso, faz-se um convite ao leitor para iniciar suas análises explorando cuidadosamente o semivariograma de nuvens que é, certamente, o "negativo" a ser revelado fielmente nos mapas de predição. Obviamente a qualidade da predição não depende somente do semivariograma de nuvens (que é o retrato fiel do processo estocástico amostrado), mas também de um bom planejamento amostral associado ao ajuste adequado do modelo teórico. Porém, perceba que o modelo é ajustado com base no gráfico do estimador de Matheron, que é uma função estatística do semivariograma de nuvens que, por sua vez, é a base de toda a fase de caracterização do processo espacial.

Devido às propriedades ótimas do semivariograma experimental de nuvens em estimar o padrão espacial, conforme abordado anteriormente, é que se

pensou o semivariograma de quantis baseado na reamostragem das similaridades estimadas, considerando as classes de distâncias h_k. Dessa forma, pode-se reproduzir fielmente o processo estocástico mediante 2.000 reamostragem, com reposição, tornando uma realidade a repetição do experimento, algo impossível nas práticas de Geoestatística convencional, para se estudar a variabilidade espacial existente na estrutura de covariância estimada. Os benefícios dessa abordagem podem ser vistos na Figura 4.8, em que cada modelo ajustado estima os parâmetros de covariância individualmente, possibilitando avaliar as incertezas associadas ao teor cobre (em ppm). Observe que esta forma de modelar a dependência espacial da variável cobre propiciou a obtenção de intervalos de confiança para os parâmetros pepita, patamar e alcance, tornando a inferência mais confiável. No estudo da variável cobre, aplicou-se a metodologia construindo o semivariograma percentil, como mostrado na Figura 4.8 e na Tabela 4.3, apenas como um exemplo geral de funcionamento do procedimento, mas quaisquer quantis podem ser obtidos, flexibilizando a construção dos intervalos de confiança de qualquer amplitude de cobertura desejada.

Na Tabela 4.3, podem ser vistos os resultados númericos do semivariograma experimental de quantis bootstrap, tendo a última coluna o viés bootstrap relativo a cada estrato ou classe de distância h. Note que os vieses foram relativamente baixos, considerando a escala dos dados, tendo que o viés médio sido 0,01750289. Essa média positiva mostra que o bootstrap baseado em $B = 2.000$ repetições por classe de distância teve uma leve tendência de superestimar as semivariâncias. Outro resultado que chama a atenção são os valores bastantes próximos dos quantis mediano $\hat{\gamma}_{50\%}$ e médio $\bar{\gamma}$ (veja Tabela 4.3). Esse fato pode ser um indício de que a distribuição das similaridades dentro de um particular estrato é razoavelmente simétrica.

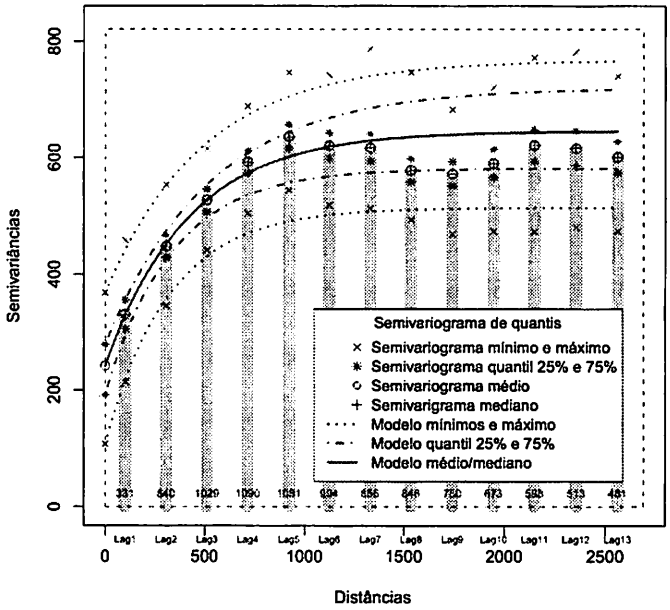


Figura 4.8 Exemplo de um semivariograma de quantis estimando o padrão de dependência espacial, da variável cobre (dados *Meuse*), por meio do modelo exponencial. Os valores na parte inferior do gráfico são os números de pares de distâncias de cada *lag*

Como se observa na Tabela 4.4, os resultados dos parâmetros estimados pelos modelos esféricos, da variável cobre, descritos na Figura 4.8, permitem uma análise mais elaborada da estrutura de covariância espacial, permitindo ao pesquisador estimar as incertezas certamente existentes na estimação dos parâmetros dos modelos. Por exemplo, em uma visão mais geral, tem-se o parâmetro φ variando entre um mínimo e um valor máximo iguais a 969,61 e 1.495,00. Outra informação importante é o quantil mediano ($\hat{\gamma}_{50\%}$) dos parâmetros estimados: entende-se que 50% das estimativas obtidas dos parâmetros pepita, patamar e alcance atingem os respectivos valores (241,78; 403,91; 1253,68). Sem perda de generalida-

Tabela 4.3 Resultado do semivariograma experimental de quantis bootstrap da variável cobre, coletada no rio *Meuse* e o viés bootstrap. A terceira coluna traz o número de pares e a útima coluna contém o viés bootstrap.

Lag	$\hat{\gamma}_{Min}$	$\hat{\gamma}_{25\%}$	$\hat{\gamma}_{50\%}$	$\bar{\gamma}$	$\hat{\gamma}_{75\%}$	$\hat{\gamma}_{Max}$	Viés.
1	213,53	304,98	329,23	331,12	355,60	458,27	0,034
2	345,78	429,37	447,74	448,24	466,86	552,56	0,31
3	440,86	506,50	527,25	526,59	545,42	615,10	0,24
4	503,90	572,51	591,20	592,18	611,77	688,67	1,05
5	543,31	615,95	634,74	636,42	656,67	746,89	-0,41
6	518,54	598,36	620,38	620,44	642,52	742,68	0,29
7	512,47	593,08	615,55	616,88	640,41	786,59	-0,90
8	493,42	557,41	576,65	577,66	597,61	747,08	0,60
9	468,60	550,32	571,06	571,80	592,75	682,80	0,39
10	473,38	565,81	589,39	589,67	614,42	720,86	-0,05
11	472,85	593,00	620,62	620,94	647,66	772,09	-0,72
12	480,21	585,82	616,00	615,84	645,77	781,61	0,10
13	473,48	573,12	600,26	600,59	627,20	740,08	-0,71

des, assume-se que esses parâmetros são também estimativas da média ($\hat{\gamma}_{50\%}$), visto que um modelo apenas foi capaz de capaz de estimar ambas as estatísticas (inspecione o gráfico da Figura 4.8). Muitas outras leituras ou conclusões podem ser tiradas com essa metodologia, conforme já é de praxe na estatística clássica. Porém, para os procedimentos de modelagem em Geoestatística, trata-se de uma abordagem inovadora, isto é, têm-se mais recursos para uma análise exploratória variográfica à disposição dos pesquisadores em Geoestatística. Note que essa metodologia vai muito mais além do que apresentado nesse simples exemplo de avaliação da variável cobre. Assim, a ideia pode ser ampliada para se construir quaisquer intervalos de confiança e medir as incertezas das estrutura de correlação da variável cobre, o que implica em predições mais precisas. Portanto, a metodologia oferece condições perfeitas para o refinamento do modelo de predição.

Tabela 4.4 Análise variográfica: ajuste do semivariograma de quantis da variável cobre, coletada no rio *Meuse*. τ^2 é a variância pepita, ϱ^2 é a contribuição e φ o alcance.

Parâmetro	$\hat{\gamma}_{Min}$	$\hat{\gamma}_{25\%}$	$\bar{\gamma} \cong \hat{\gamma}_{50\%}$	$\hat{\gamma}_{75\%}$	$\hat{\gamma}_{Max}$
τ^2	108,70	192,21	241,78	264,59	368,44
ϱ^2	405,13	388,71	403,91	440,00	399,85
φ	969,61	1006,43	1253,68	1445,73	1495,00

4.2 Teste de normalidade para Geoestatística

O Teste Gaussiano para Geoestatística (TGGeo) proposto, foi submetido à prova por meio de simulações de processos gaussianos e não gaussianos. Depois, o TGGeo foi submetido a três conjuntos de dados para se avaliar o seu desempenho em situações do mundo real, em que as condições não são controladas e a distribuição dos dados é totalmente desconhecida. Estes são processos estocásticos ocorrendo na natureza, a serem explorados pela primeira vez nesse contexto. Aliás, o TGGeo é o primeiro teste proposto, até o momento, na área de Geoestatística, que se utiliza de artefatos produzidos pela própria Geoestatística. A seguir, esses resultados são particularmente apresentados e discutidos ao seu tempo.

4.2.1 Campos aleatórios gaussianos sob correlação esférica

Conforme pode ser visto na Figura 4.9, o TGGeo reconhece a normalidade multivariada presente nos dados previamente simulados, no contexto do cenário I, considerando a estrutura $\{\mu = 3{,}0; \tau^2 = 0{,}5; \sigma^2 = 5{,}0; \varphi = 2{,}0; n = 121; \text{modelo} = sph\}$. Após estimar o semivariograma experimental, procedeu-se o ajuste por meio modelo esférico (sph), via critério de MV e do QMP, os quais modelaram os dados de forma semelhante, a julgar pelos limites de confi-

ança percentil bootstrap, gerado por $B = 300$ iterações. Portanto, com base no valor-pl = 1,0000, aceita-se a hipótese H_0 de os dados terem distribuição gaussiana.

O resultado obtido é de grande relevância tanto para a Geoestatística convencional como para a Geoestatística baseada em modelo. O fato de ser comprovado que o processo estocástico é gaussiano, possibilita a construção de intervalos de confiança para $\hat{\gamma}(h)$, aproximado pela distribibuição Qui-quadrado (χ^2_1), com 1 grau de liberdade, conforme já se utiliza na Estatística em geral (GUMPERTZ, 1999; CRESSIE, 1993).

4.2.2 Campos aleatórios gaussianos sob correlação exponencial

Dado que foram simuladas as variáveis estacionárias gaussianas, sem tendência e sob condições isotrópicas, o TGGeo dá indícios da presença de normalidade multivariada, conforme mostrado na Figura 4.10. Em termos da MV, a curva do modelo exponencial (exp) teve apenas 2,79% das unidades preditivas (de um total de 2.000 unidades) fora das bandas de confiança bootstrap, mantendo-se conservadora em relação ao valor nominal prefixado. Então, com em um intervalo de 95% de confiança, o teste não rejeitou a hipótese de nulidade, como indicado pela estatística valor-pl = 0,968333 (Figura 4.10). Claramente, os dois critérios de suavização do modelo exponencial, para esse caso, estimaram satisfatoriamente os parâmetros do mesmo processo estocástico (mesma população). Em outras palavras, o TGGeo mostrou-se capaz de comprovar a existência de processo gaussiano nos dados simulados, sob correlação (espacial) exponencial do cenário I.

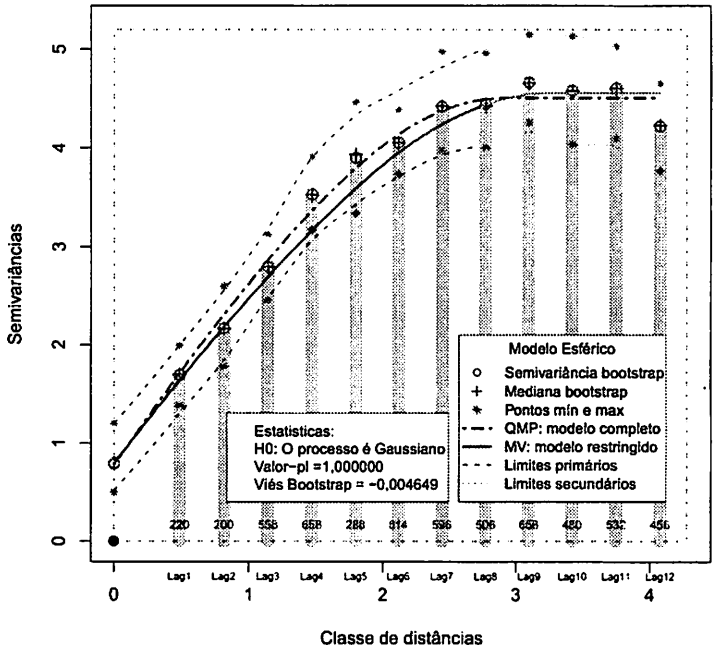


Figura 4.9 O TGGeo acusa normalidade multivariada nos dados, sendo os dados simulados assumindo distribuição Gaussiana, com tendência direcional constante, fixando os seguintes parâmetros: $\{\mu = 3{,}0; \tau^2 = 0{,}5; \varrho^2 = 4{,}5; \varphi = 2{,}00; n = 121;$ modelo=sph$\}$ – a simulação foi realizada pela função *GausRF*

4.2.3 Campos aleatórios gaussianos sob correlação gaussiana

Apresenta-se uma terceira e última simulação do cenário I para avaliar o TGGeo, supondo uma estrutura de dependência espacial gaussiana (gau), gerada por um processo estocástico gaussiano, definido pelos parâmetros $\{\mu = 4{,}2; \tau^2 = 1{,}2; \sigma^2 = 11{,}2; \varphi = 2{,}2; n = 100;$ modelo=gau$\}$. O resultado do teste está representado na Figura 4.11 e, igualmente, nos casos anteriormente apresentados, o TGGeo mostou-se eficiente para evidenciar a existência de um processo gaussiano

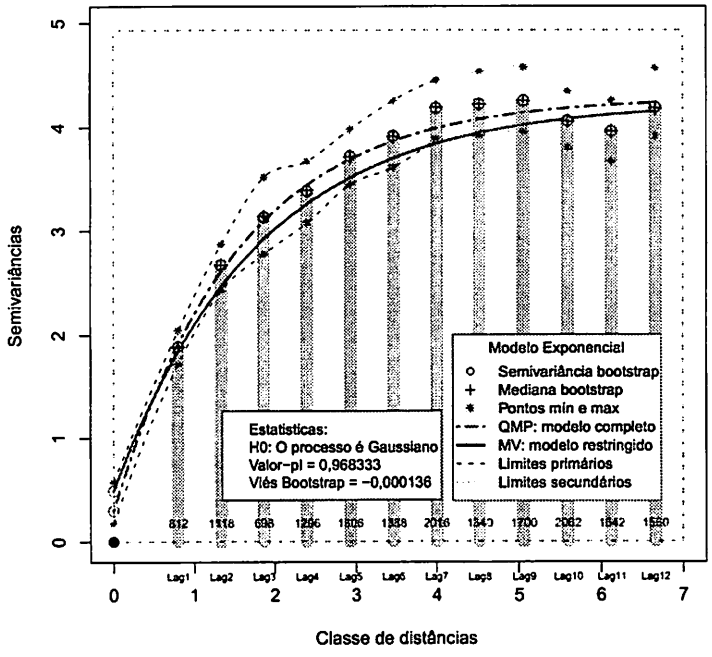


Figura 4.10 TGGeo identifica o campo aleatório gaussiano, simulado segundo o parâmetros $\{\mu = 0; \tau^2 = 0{,}7; \sigma^2 = 6{,}7; \varphi = 3{,}5; n = 225;$ modelo=exp$\}$, sob a usando a função *GaussRF* e intervalo de confiança percentil bootstrap de 95%

estabelecido por simulação. Na Figura 4.11 observa-se claramente que não há diferença entre os modelos ajustados pelo critério da MV e o dos QMP, garantido pelo intervalo de confiança percentil de 95% e comprovado pelo valor-pl = 1,0000. A boa performance do método da MV não está relacionada com bom comportamento da nuvem de pontos do semivariograma experimental, mas porque esse critério é adequado para estimar parâmetros de uma processo aleatório gaussiano, independente do tipo de padrão espacial. Nesses condições, espara-se que ambos os métodos, MV e QM, gerem as mesmas estimativas dos parâmetros de covariân-

cia.

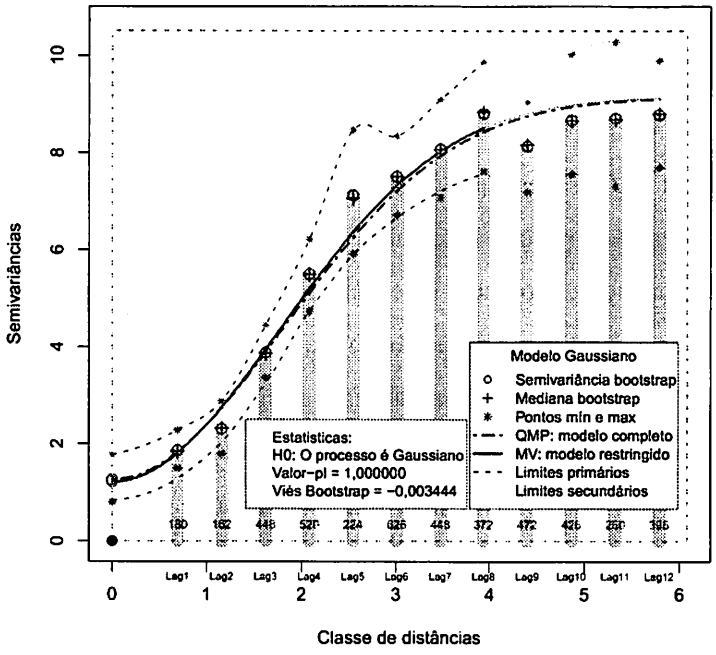


Figura 4.11 Representação gráfico do TGGeo, ratificando a existância de um processo estocástico simulado por meio da função *GaussRF*, considerando os seguintes parâmetros $\{\mu = 4{,}2; \tau^2 = 1{,}2; \sigma^2 = 11{,}2; \varphi = 2{,}2; n = 100;$ modelo=gau$\}$ e intervalo de confiança percentil bootstrap de 95%

A performance (ou eficiência) dos critérios dos QMP e da MV para estimar os parâmetros de covariância, em todos os casos do cenário I, foi bastante similar. Esse fato é ratificado por meio dos resultados apresentados na Tabela 4.5, os quais mostram que os parâmetros estimados por esses tais critério são estatisticamente iguais. Observe que a estimação dos parâmetros de cada um dos três modelos considerados não foi pontual e, portanto, a construção dos intervalos de confiança percentis bootstrap proporcionou uma margem de segurança, com 95%

de confiança, para o analista comparar ou avaliar os parâmetros estimados nos dois critérios de ajuste adotado. É importante observar que uma análise descritiva desse gênero não é usual no contexto da Geoestatística convencional, devido a não possibilidade da repetição do experimento, tão comum na estatística experimental, uma vez que é impossível se aplicar o princípio da aleatorização e, muito menos, o da repetição.

É importante ressaltar que o principal interesse da Estatística espacial é o de modelar a autocorrelação espacial e, no entanto, não faz nenhum sentido pensar em aleatorização. Contudo, é de suma importância criar mecanismos que viabilizem a "repetição" do experimento, para que se possam avaliar as incertezas presentes no processo de modelagem do semivariograma, bem como na predição espacial como um todo e, com essa finalidade, sugerem-se as técnicas reamostragem bootstrap como uma possível, plausível e interessante solução.

Tabela 4.5 Resumo estatístico do cenário I.

Modelo esférico				
Parâmetro	QMP	MV	LI(2,5%)	LS(97,5%)
$\hat{\tau}^2$	0,8019	0,7988	0,4221	1,1016
$\hat{\sigma}^2$	4,5899	4,5564	3,9694	4,9903
$\hat{\varphi}$	3,0151	3,2570	2,8361	3,2137
Modelo exponencial				
Parâmetro	QMP	MV	LI(2,5%)	LS(97,5%)
$\hat{\tau}^2$	0,2910	0,4894	0,1510	0,5410
$\hat{\sigma}^2$	4,2035	4,2419	3,6552	4,5579
$\hat{\varphi}$	4,2884	5,3601	3,9379	4,4474
Modelo gaussiano				
Parâmetro	QMP	MV	LI(2,5%)	LS(97,5%)
$\hat{\tau}^2$	1,1973	1,2160	0,7116	1,6660
$\hat{\sigma}^2$	9,3948	4,5564	7,7677	4,9903
$\hat{\varphi}$	4,0962	9,1780	3,92051	9,9673

4.2.4 Campos aleatórios não gaussianos simulados sob correlação esférica

Com o mesmo propósito de testar o desempenho o TGGeo, também foram simulados três diferentes processos estocásticos espacialmente contínuos, cuja distribuição é desconhecida, os quais constam do cenário II. Para o primeiro modelo simulado, gerou-se uma amostra com a seguinte configuração: $\{\mu = 6{,}0; \tau^2 = 1{,}0; \sigma^2 = 10{,}0; \varphi = 3{,}0; n = 121;$ modelo=sph$\}$, sendo o campo aleatório obtido por uma combinação da perturbação do processo gaussiano, dado por esses argumentos fixados, com um processo de exponencial, tal que $exp(n = 121, rate =$

3,0). Sendo assim, o teste foi aplicado e, como pode ser observado na Figura 4.12, o critério da MV apresentou, aproximadamente, 68,56% das unidades preditas fora dos limites de confiança, isto é, um valor de plausibilidade valor-pl = 0,314423. Este é um caso típico em que a decisão de se rejeitar ou aceitar H_0 está firmada na qualidade do ajuste do critério da MV e não somente com base no valor-pl. Observe que esse critério estimou um valor muito baixo para o patamar. O valor obtido pela MV é completamente irreal quando comparado com o patamar do semivariograma experimental. Com 31,44% da curva do critério da MV fora do limite de confiança bootstrap, não há razões para se crer em normalidade multivariada nesse caso. Deve-se lembar que o critério da MV é apropriado para modelar campos aleatórios gaussianos (GUEDES, 2008) e, por isso, este não teve um bom desempenho, mostrando-se coerente com os resultados simulados com esse propósito. Por conseguinte, rejeita-se a hipótese de o processo ser gaussiano, com base nas evidências apresentados pelo TGGeo.

4.2.5 Campos aleatórios não gaussianos simulados sob correlação exponencial

Ainda, no intuito de avaliar a robustez do teste, simulou-se um segundo campo aleatório gaussiano, $\mathcal{R}$, perturbado também com uma distribuição exponencial, conforme descrito na metodologia. E, mais uma vez, o TGGeo foi robusto, também rejeitando a hipótese H_0 de o campo aleatório ser gaussiano, com um valor-pl = 0,194118, conforme pode ser visto na Figura 4.13. Pelas mesmas razões abordadas na seção anterior, parece prudente determinar que o processo não é gaussiano, considerando que 80,59% da curva da MV (modelo restringido) figurou fora dos limites de confiança percentil bootstrap.

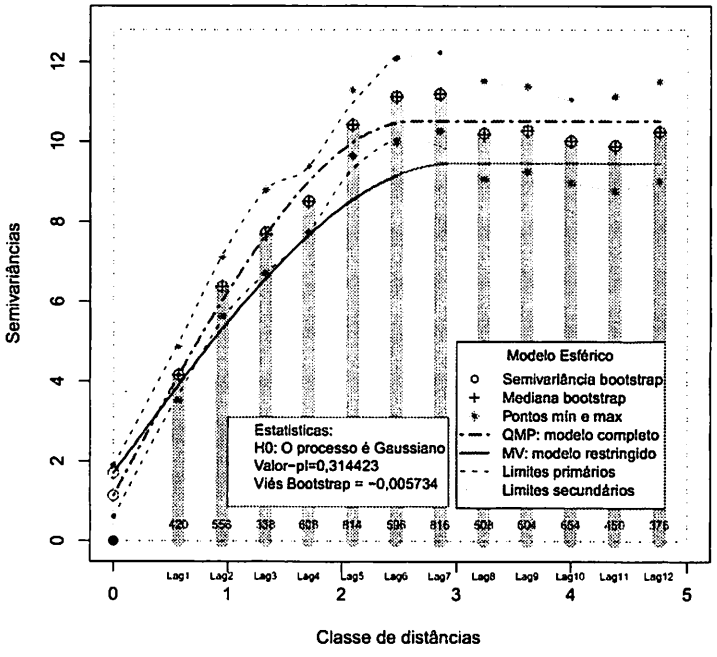


Figura 4.12 Apresentação dos resultados do TGGeo, concernente ao campo aleatório não gaussiano, simulado com uma estrutura de autocorrelação espacial esférica, fixando os argumentos, $\{\mu = 6{,}0; \tau^2 = 1{,}0; \sigma^2 = 10{,}0; \varphi = 3{,}0; n = 121;$ modelo=sph$\}$, sendo o processo gaussiano perturbado com uma exponencial, tal que exp$(n = 121, rate = 3{,}0)$. Foram realizadas $B = 500$ repetições bootstrap para gerar o intervalo de confiança percentil de 95%

4.2.6 Campos aleatórios não gaussianos simulados sob correlação gaussiana

Outra situação simulada no cenário II é apresentada, em que o processo gaussiano gerado, com base nos parâmetros $\{\mu = 6{,}0; \tau^2 = 1{,}0; \sigma^2 = 10{,}0; \varphi = 3{,}0; n = 121;$ modelo=gau$\}$, foi perturbado por uma distribuição exponencial, tal que exp$(n = 121, rate = 3{,}0)$. Após avaliar o gráfico da Figura 4.14, observa-

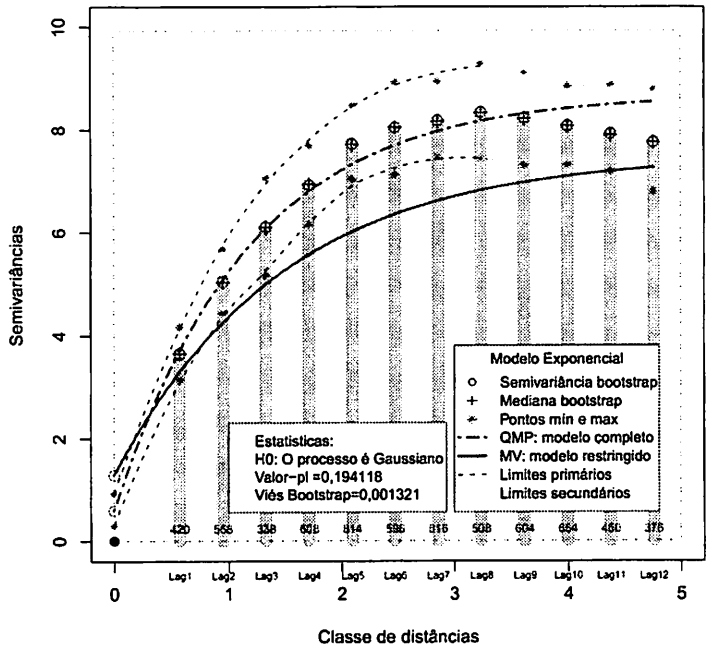


Figura 4.13 Gráfico desmostrativo do TGGeo, indicando que o processo não é gaussiano. A simulação foi configurada com os parâmetros $\{\mu = 6{,}0; \tau^2 = 1{,}0; \sigma^2 = 10{,}0; \varphi = 3{,}0; n = 121;$ modelo=exp$\}$, perturbando o processo gaussiano com uma distribuição exponencial, tal que exp($n = 121$,$rate = 3{,}0$). Utilizou-se $B = 500$ repetições bootstrap para gerar o intervalo de confiança percentil de 95%

se que TGGeo mostrou ser novamente robusto e de forma consistente rejeitou a hipótese H_0: os dados são Gaussianos. Note que o modelo da MV teve 75,43% da curva fora dos limites com 95% de confiança, indicando que o teste teve uma razoável plausibilidade ao rejeitar H_0, sendo que o processo não é gaussiano, isto é, H_0 não é verdadeira (o processo não é gaussiano).

O teste também revelou alguns detalhes no cenário II que merecem ser salientados. Em primeiro lugar, visivelmente, percebe-se que o critério da máxima

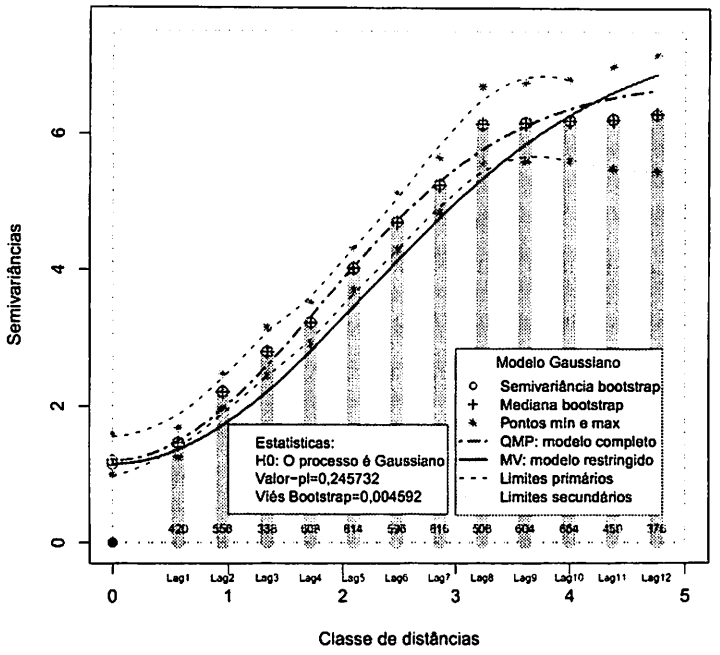


Figura 4.14 Apresentação dos resultados do TGGeo, considerando o processso estocástico não gaussiano, obtido a partir de um processo gaussiano simulado com os parâmetros $\{\mu = 6{,}0; \tau^2 = 1{,}0; \sigma^2 = 10{,}0; \varphi = 3{,}0; n = 121;$ modelo=exp$\}$ perturbado com uma exponencial, tal que $\exp(n = 121, rate = 3{,}0)$. O intervalo de confiança percentil boostrap foi construído com $B = 500$ repetições

verossimilhaça apresenta inadequações para o ajuste do semivariograma. De modo geral, nos dois primeiros casos, para esse critério (MV), o modelo teve um bom comportamento no início do semivariograma, mas não estimou bem o patamar, em nenhuma das situações aqui avaliadas, tendo uma clara tendência de subestimá-lo, como pode ser constatado nas Figuras 4.12, 4.13 e 4.14. Já o efeito pepita tendeu a ser superestimado, exceto o modelo sob correlação guassiana. Isso pode estar relacionado ao fato de a MV desconsiderar a nuvem de pontos do semivariograma

e ajustar o modelo com base nos dados. Esse fato torna esse critério de ajuste altamente sensível ao comportamento do campo aleatório (processo estocástico) e, por isso, o mesmo deve ser usado com cautela. Essa recomendação está respaldada na exigência de normalidade multivariada requerida por esse critério, sem a qual os resultados não são confiáveis. Por outro lado, o método dos quadrados mínimos não é afetado pelo processo estocástico e não exige nenhum tipo de pressuposição. Assim, o leitor acostumado com os clássicos ajustes de regressão e iniciante em Geoestatística deve estranhar um pouco o comportamanto do critério da MV. Esse comportamento pode, talvez, conduzir ao equívoco de se pensar que o modelo baseado em MV está estimando os parâmetros de um universo totalmente diferente que o método dos QM, o que não é verdade. De fato, trata-se de uma demostração bastante coerente com a natureza dos dados simulados no cenário II, uma vez que, na falta total de normalidade multivariada do campo aleatório, os dados não são oriundos de um processo estocástico gaussiano. No entanto, uma importante conclusão pode ser tirada: para esse tipo de cenário, o critério da máxima verossimilhança não se mostrou adequado para se caracterizar o padrão de dependência do fenômeno. Esse raciocínio lógico justifica os fundamentos do TGGeo e também o qualifica como uma ferramenta exclusiva, em aprimoramento, para investigar a presença de processos estocásticos gaussianos.

Assim, pode-se conceber que o Teste Gaussiano para Geoestatística tem uma filosofia bastante simples:

Axioma: *Em geral, se as variáveis regionalizadas $Z(x)$ são realizações de um processo estocástico gaussiano (espacialmente contínuos evidentemente), então, espera-se, com $\delta \times 100\%$ de chance, mediante as técnicas de reamostragem bootstrap, que o critério da MV seja estatisticamente idêntico ao método dos quadrados mínimos.*

Esse achado, certamente, promove a Geoestatística baseada em modelos, dando-lhe um caráter de aplicação no mundo real.

A seguir, serão apresentados os resultados de aplicação do TGGeo em dados reais.

4.2.7 Teste gaussiano para Geoestatística: uma aplicação em dados reais

A julgar pelo gráfico da Figura 4.15, o TGGeo acusou que o conjunto de dados de argila é uma possível amostra de um processo estocástico gaussiano subjacente, apresentando baixa evidência contra H_0 e o modelo selecionado foi a estrutura de autorrelação esférica. Verifica-se que ambas as linhas preditivas modelaram de modo satisfatório o teor de argila e podem ser usadas para fins de predição espacial, com um valor-pl = 0,208088. Perceba que, para se processar o TGGEo, é necessario realizar todo o procedimento padrão da análise variográfica. Assim, os parâmetros de covariância espacial foram estimados pelos critérios da MV e dos QMP, de forma que um pequeno resumo estatístico foi obtido e está apresentado na Tabela 4.6.

Verdadeiramente, esses resultados sinalizam que a amostra vem de uma população com distribuição normal multivariada, o que é um importante resultado encontrado. Por outro lado, todos os parâmetros de covariância caíram dentro do intervalo percentil bootstrap de 95% de confiança, indicando que não existe diferença significativa entre os ajustes. Esse fato implica dizer que o estimador de máxima verossimilhança foi equivalente ao estimador de quadrados mínimos ponderado, para avaliar o padrão de dependência espacial da variável argila.

O TGGeo também permite uma análise subjetivamente visual, a julgar pela qualidade dos ajustes feitos no semivariograma experimental, constatada pela

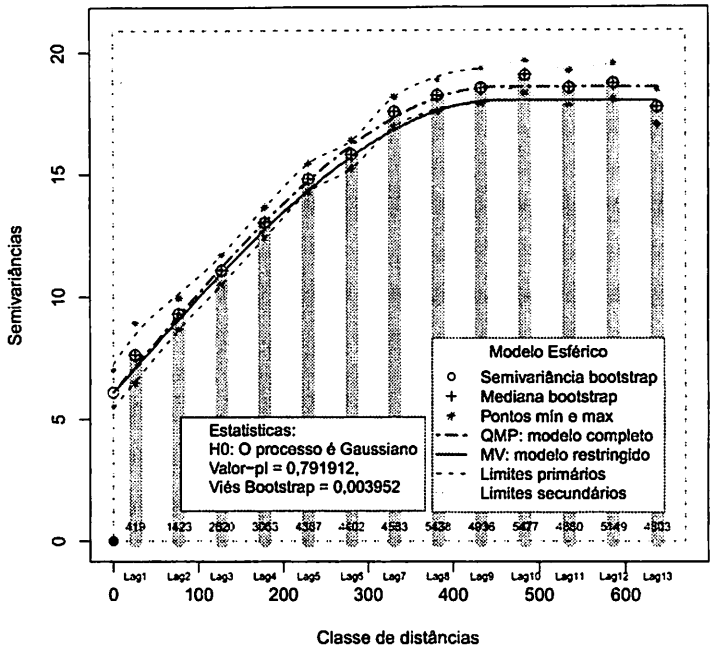


Figura 4.15 Resultado do TGGeo aplicado nos dados de argila
Fonte: os dados foram cedidos por Sandro Manoel C. Hurtado (HURTADO, 2008)

razoável proximidade das curvas inseridas nas bandas de confiança bootstrap de 95%. Observe que não se trata de uma comparação de modelos, mas de um critério que proveu substancial recurso para se tomar a decisão de aceitar a hipótese, H_0, de a distribuição da variável argila ser gaussiana. E, de fato, há forte indício de esse processo estocástico ser gaussiano.

A Figura 4.16 é referente ao resultado do TGGeo aplicado à variável matéria orgânica, dos dados rio *Meuse* (PEBESMA, 2004). Nessa avaliação, o teste não apresenta indícios de o processo ser gaussiano, rejeitando a hipótese de H_0, devido ao fato de 67,14% das unidades preditivas do critério da MV estarem fora do intervalo de confiança percentil bootstrap. Colocando de outra forma, 67,14%

Tabela 4.6 Análise variográfica e intervalo de confiança percentil bootstrap, dos dados de argila.

Parâmetro	QMP	MV	LI(2,5%)	LS(97.5%)
$\hat{\tau}^2$	6,10	6,07	5,50	7,00
$\hat{\sigma}^2$	18,59	18,04	16,89	19,09
$\hat{\varphi}$	455,00	454,99	412,00	455,00

dos valores preditos pelo critério da MV não pertencem ao intervalo de confiança contra 100% das unidades preditivas do método dos QMP que estão dentro desses limites. Obervando-se a Figura 4.16, é claramente visível a falta de ajuste do critério da MV que, nem de longe, se aproxima das estimativas dos parâmetros de covariância obtidos pelo método dos QMP, que modela perfeitamente a variável matéria orgânica. A única explicação lógica para esse fato é que, realmente, a distribuição dos dados matéria orgânica não possui distribuição normal multivariada. Portanto, o TGGeo apresentou um resultado significativo quanto à natureza desses dados avaliados, indicando falta de normalidade do processo com base no péssimo desempenho do modelo ajustado pela MV.

Assim, a amostra da variável matéria orgânica possui uma distribuição diferente da gaussiana, o que impossibilita o ajuste por meio do critério da MV, tornando-o inadequado e, portanto, o método dos QMP deve ser preferido. Finalmente, o TGGeo foi aplicado a dados de cafeeiros amostrados no estado de Minas Gerais, em 2005. Como pode ser visto na Figura 4.17, o teste indicou que o processo segue uma distribuição gaussiana, com um valor-pl = 1,0000, tendo realizado $B = 1.000$ repetições bootstrap. Embora, a curva de MV, visivelmente, não tenha ficado muito próxima da curva dos QMP, não extrapolou os limites de confiança percentil, ratificando o indício de que os dados da amostra de café devem ser realizações de um campo aleatório gaussiano. Esse resultado é muito importante para os pesquisadores de café que estudam a dinâmica espaço-temporal dessa cul-

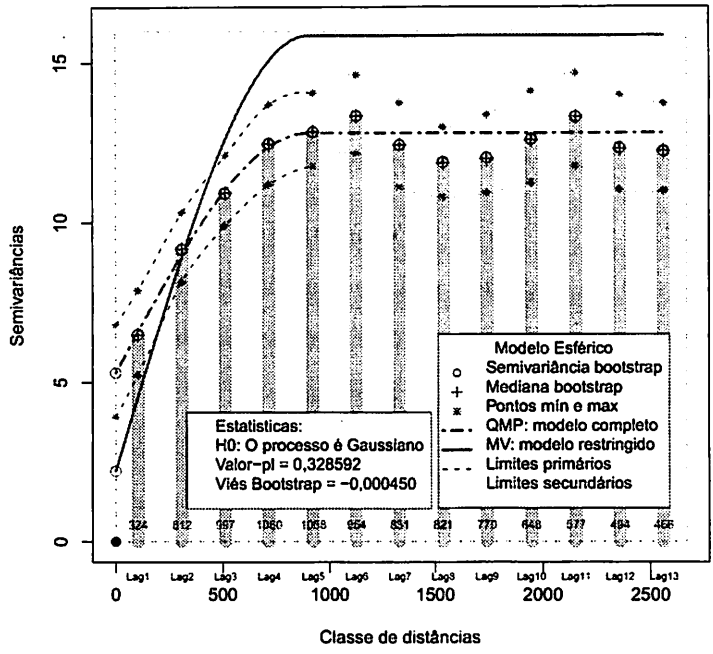


Figura 4.16 Resultado do TGGeo aplicado à variável matéria orgânica, dos dados do rio *Meuse*, coletado nas proximidades da vila de Stein, Utrecht, 1993

tura, tendo em vista a facilidade de se fazer inferência utilizando as já consolidadas ferramentas da Estatística clássica, tais como: comparar médias espaciais por testes, realizar intervalos de confiança para o semivariograma usando a aproximação pela distribuição Qui-quadrado, realizar teste de hipóteses, etc. O leitor deve estar ciente da sutileza existente quanto ao uso indiscriminado do critério da máxima verossimilhança na modelagem do semivariograma. Como ocorre na Estatística clássica, o critério da MV só pode modelar aqueles fenômenos que certamente são realizações de processo estocástico gausssiano. Caso contrário, as estimativas encontradas estão erradas e, consequentemente, a predição espacial, se for objeto

de interesse, está totalmente equivocada. Essa sutileza, que não pode ser negligenciada, até então, inviabilizava o uso prático do critério da MV em qualquer modelagem em Geoestatística, uma vez que não era possível verificar se o processo é gaussiano, ou não, a princípio. Então, as análises eram realizadas por meio de suposições teóricas, sem a devida confirmação, o que agora é possível com a proposição do TGGeo.

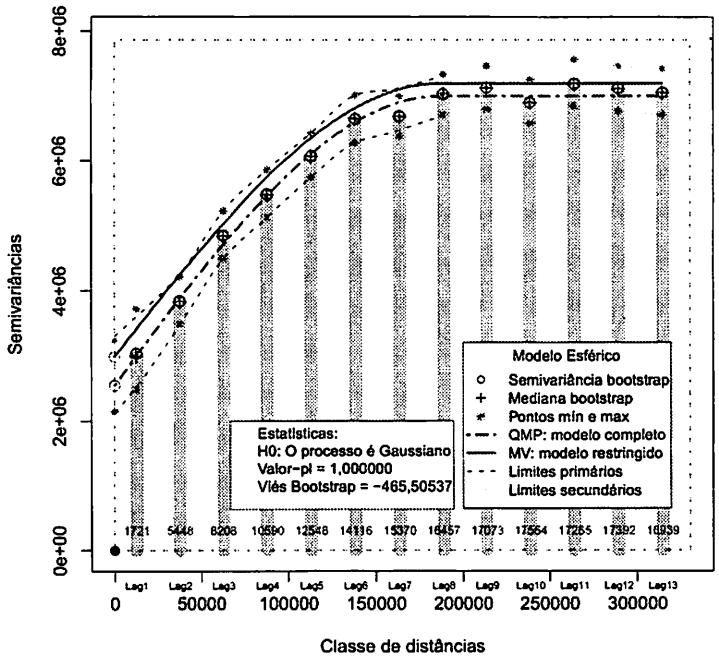


Figura 4.17 Resultado do teste normalidade dos dados de cafeeiro (*Coffea arabica* L e *Coffea sp.*), em toneladas, coletados pelo IBGE, no estado de Minas Gerais, em 2005, considerando-se a sede municipal como unidade de amostragem

Na Tabela 4.7 é apresentado o resultado da modelagem dos semivariogramas dos dados de cafeeiro, bem como o intervalo de confiança, para que se possa avaliar o

desempenho dos modelos ajustados e, também, as incertezas associados aos parâmetros de covariância estimados pela MV e pelos métodos dos QMP.

Tabela 4.7 Análise variográfica e intervalo de confiança percentil bootstrap dos dados de cafeeiro, do estado de Minas Gerais, coletados em 2005.

Parâmetro	QMP	MV	LI(2,5%)	LS(97,5%)
$\hat{\tau}^2$	2.545.567,00	2.981.915,00	2.146.456,20	3.236.521,20
$\hat{\sigma}^2$	6.990.334,00	7.186.298,00	6.593.391,40	738.5240,40
$\hat{\varphi}$	186.235,00	186.235,00	181.399,60	201.982,40

5 CONCLUSÕES

A metodologia bootstrap foi adequada para avaliar as incertezas associadas ao processo de estimação do semivariograma experimental, por possibilitar a construção de intervalos de confiança percentil, a partir do semivariograma de nuvens.

Foi adequada e essencial a utilização do semivariograma de nuvens (semivariograma *cloud*) como base de pseudodados para se processar as devidas repetições boostrap, possibilitando a construção do TGGeo e do semivariograma de quantis.

O TGGeo é uma resposta interessante ao problema de pesquisa proposto nesta tese. Ele viabiliza razoável instrumento para se testar a normalidade multivariada em dados continuamente correlacionados no espaço, algo inédito em Geoestatística. Sua importância é ratificada pelos resultados obtidos nos três cenários avaliados, em que o desempenho do teste foi avaliado com dados reais e, por meio de simulações, o TGGeo mostrou-se apto para identificar processos espaciais gaussianos, bem como os não gaussianos. Esse conhecimento é de relevante importância científica.

Os resultados obtidos para as simulações dos cenários I e II mostraram que o TGGeo apresentou boa performance: nos processos gaussianos simulados (cenário I), o TGGeo confirmou a presença de normalidade e, em relação aos processos não gaussianos simulados (cenário II), mostrou-se robusto, rejeitando a suposição de o processo ser gaussiano.

O critério de ponderação $W_{N/h}$, que atribui maiores pesos aos *lags* iniciais e penaliza, exponencialmente decrescente, com menor peso os *lags* depois do alcance, parece corrigir possíveis problemas de menores número de pares dos pri-

meiros *lags*, o que geralmente é encontrados em "lattices" irregulares. Contudo, sua eficiência deve ser melhor estudada.

REFERÊNCIAS

ARBIA, G. **Statistical foundation and applications to regional convergence.** Berlin: Springer, 2006. 207 p.

AZZALINI, A. **Statistical inference**: based on Likelihood. London: Chapman and Hall, 1996.

BABU, G. J. Subsample and half-sample methods. **Annals of the Institute of Statistical Mathematics**, Tokyo, v. 44, n. 4, p. 703-720, Dec. 1992.

BATES, C.; WHITE, H. A unified theory of consistent estimation for parametric models. **Econometric Theory, Cambridge**, v. 1, n. 2, p. 151-178, Aug. 1985.

BICKEL , J. P.; FREEDMAN, D. A. Some asymptotic theory for the bootstrap. **Annals of Statistics, Hayward**, v. 9, n. 6, p. 1196-1217, Nov. 1981.

BOX, G. E. P.; COX, D. R. An analysis of transformation.**Journal of the Royal Statistical Society**: serie B, methodological, London, v. 26, n. 2, p. 211-243, July 1964.

CAMBARDELLA, C. A. et al. Field-scale variability of soil properties in Central Iowa soils. **Soil Science Society of America Journal**, Madison, v. 58, n. 5, p. 1501-1511, Sept./ Oct. 1994.

CHERNICK, M. R. **Bootstrap Methods**: a guide for practitioners and researchers. 2th ed. New Jersey: J. Wiley, 2008.

CHILÈS, J. P.; DELFINER, P. **Geostatistics**: modeling spatial uncertainty. New York: J. Wiley, 1999. 695 p.

CHRISTENSEN, O.; RIBEIRO JUNIOR, P. georglm - a package for generalised linear spatial models. **R-News**, v. 2, n. 2, p. 26-28, 2002. Disponível em: <http://cran.R-project.org/doc/Rnews>. Acesso em: 18 jan. 2011.

COCHRAN, W. G. **Sampling techniques**. New York: J. Wiley, 1953.

CRAWLEY, M. J. **Statistical computing**: an introduction to data analysis using s-plus. London: J. Wiley, 2006. 761 p.

CRESSIE, N. Fitting variogram models by weighted least squares. **Mathematical Geology**, New York, v. 17, n. 5, p. 563-585, July 1985.

CRESSIE, N. The origins of kriging. **Mathematical Geology**, New York, v. 22, n. 3, p. 239-252, Sept. 1989.

CRESSIE, N. A. **Statistics for spatial data**. New York: Wiley-Interscience, 1993.

DAVISON, A. C.; HINKLEY, D. V. **Bootstrap methods and their application**. Cambridge: Cambridge, 1997.

DIGGLE, P. J. et al. Childhood malaria in the gambia: a case-study in model-based geostatistics. **Journal of the Royal Statistical Society**: series C, applied statistics, London, v. 51, n. 4, p. 493-506, Sept. 2002.

DIGGLE, P. J.; RIBEIRO JUNIOR, P. J. **Model-based geostatistics**. New York: Springer, 2007.

DIGGLE, P. J.; RIBEIRO JUNIOR, P. J.; CHRISTENSEN, O. F. **An introdution to model–based geostatistics**. New York: Springer, 2003.

DIGGLE, P. J.; TAWN, J. A.; MOYEED, R. A. Model based geostatistics (with discussion). **Journal of the Royal Statistical Society**: series C, applied statistics, London, v. 47, n.3, p. 299-350, 1998.

EFRON, B. Bootstrap methods; another look at the jackknife. **Annals of Statistics**, Hayward, v. 7, n. 1, p. 1-26, Jan. 1979.

EFRON, B.; TIBSHIRANI, R. J. **An Introduction to the boostrap**. New York: Chapman & Hall, 1993. 430 p.

FALK, M.; KAUFMANN, E. Coverage probabilities of bootstrap-confidence intervals for quantiles. **Annals of Statistics**, Hayward, v. 19, n. 1, p. 485-495, Mar. 1991.

GOOD, P. I. **Permatation, Parametric and Bootstrap Test of Hypotheses**. 3th ed. New York: Springer, 2005. 315 p.

GUEDES, L. P. C. **Otimização de amostragem espacial**. 2008. 143 f. Tese (Doutorado em Agronomia) - Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2008.

GUMPERTZ, M. **Applied spatial statisticas**. North Carolina: NCSU, 1999.

Disponível em: <http://www/stat.ncsu.edu/classes/st564/>. Acesso em: 7 set. 1999.

HARTIGAN, J. A. Using subsample values as typical values. **Journal of the American Statistical Association**, New York, v. 64, n. 328, p. 1303-1317, Dec. 1969.

HARTIGAN, J. A. Error analysis by replaced samples. **Journal of the Royal Statistical Society**: serie B, methodological, London, v. 33, n. 1, p. 1098-1130, 1971.

HARTIGAN, J. A. Necessary and suffi cient conditions for asymptotic joint normality of a statistic and its subsample values. **Annals of Statistics**, Hayward, v. 3, n. 3, p. 573-580, May 1975.

HARVILLE, D. A. Bayesian inference for variance components using only error contrasts. **Biometrika**, London, n. 61, n. 2, p. 383-385, Aug. 1974.

HELMERS, R.; JANSSEN, P.; VERAVERBEKE, N. Bootstrapping u-quantiles. In: LEPAGE, R.; BILLARD, L. (Ed.). **Exploring the limits of bootstrap**. New York: J. Wiley, 1992. p. 145-155.

HIJMANS, R.; MAGNUS, J. Consistent maximum likelihood estimation with dependent observations: the general (non-normal) case and the normal case. **Journal of Econometrics**, Amsterdam, v. 32, n. 2, p. 253-285, July 1986.

HURTADO, S. M. C. **Uso do clorofilômetro e de agricultura de precisão no manejo da adubação nitrogenada do milho**. 2008. 92 f. Tese (Doutorado em Agronomia) - Universidade Federal de Lavras, Lavras, 2008.

ISAAKS, E. H.; SRIVASTAVA, R. M. **Applied geostatistics**. New York: Oxford University, 1989.

JOURNEL, A. G.; HUIJBREGTS, C. J. **Mining Geostatisitcs**. 5th ed. London: Academic, 1991. 600p.

JUMARS, P. A.; THLSTLE, D.; JONES, M. L. Detecting two-dmensional spatial structure in biological data. **Oecologia**, Berlin, v. 28, n. 2, p. 109-123, June 1977.

KANNAN, D. **An introduction to stochastic processes**. Limerick: Elsevier, 1979. 296 p.

KRIGE, D. G. A statistical approach to some basic mine valuation problems on

the witwatersrand. **Journal of the Chemical, Metallurgical & Mining Society of South Africa**, Johannesburg, v. 52, n. 6, p. 119-139, Dec. 1951.

KRIGE, D. G. A pratical review of some basic concepts and tchniques for mining application. In: INTERNATIONAL SYMPOSIUM ON APPLICATION OF COMPUTERS AND OPERATIONS RESEARCH IN THE MINERAL INDUSTRY, 33., 2007, Santiago. **Anais**... Santiago: APCOM, 2007. Disponível em: <http://www.apcom2007.com>. Acesso em: 16 out. 2009.

LEHMANN, E. L. **Elements of large-sample theory**. New York: Springer-Verlag, 1999.

LEHMANN, E. L.; CASELLA, G. E. **Theory of point estimation**. 2th ed. New York: Springer-Verlag, 1998. 153 p.

LITTEL, R. C. et al. **SAS system for mixed models**. North Carolina: SAS, 1996. 633 p.

MANLY, B. F. J. **Randomization and monte carlo methods in biogoly**. London: Chapman, 1991.

MATÉRN, B. Spatial variation: stochastic models and their application to some problems in forest surveys and other sampling investigations. **Meddelanden Fran Statens Skogsforskningsinstitut**, Stockholm, v. 49, n. 5, p. 114, 1960.

MATHERON, G. **Traite de geostatistique appliquee**. Paris: Technip, 1962.

MATHERON, G. **The theory of regionalized variables and its applications**. Paris: Ecole Nationale Supe Rieuredes Mines, 1971.

MCCARTHY, P. J. Pseudo-replication: half-samples. **International Statistical Review**, Edinburgh, v. 37, n. 3, p. 239-263, 1969.

MOORE, M. **Bootstrap Methods and Perutation (wiht notes)**. 5th ed. New York: W. H. Freeman, 2006.

MÜLLER, W. G. **Collecting spatial data**: optimum design of experiments for random fields. 3th ed. Berlin: Springer-Verlag, 2007.

OLIVEIRA, M. S. de. **Plano amostral para variáveis espaciais utilizando geoestatística**. 1991. 100 f. Dissertação (Mestrado em Estatística) - Universidade de Campinas, Campinas, 1991.

PATTERSON, H. D.; THOMPSON, R. Recovery of interblock information when

block sizes are unequal. **Biomatrika**, London, v. 58, n. 3, p. 545-554, Dec. 1971.

PEBESMA, E. J. Multivariable geostatistics in s: the gstat package. **Computers and Geosciences**, New York, v. 30, n. 7, p. 683-691, Aug. 2004.

R DEVELOPMENT CORE TEAM. **R**: a language and environment for statistical computing. Vienna, 2008. Disponível em: <http://www.R-project.org>. Acesso em: 16 out. 2009.

RAO, C. R. Minqe theory and its relation to ml and mml estimation of variance components. **Sankhyā**: the indian journal of statistics, series b, Calcutta, n. 41, n. 3/4, p. 31-49, Dec. 1979.

RIBEIRO JUNIOR, P. J.; DIGGLE, P. J. GeoR: a package for geostatistical analysis. **Analysis**, Oxford, v. 1, n. 2, p. 14-18, June 2001.

RIZZO, M. L. **Statistical computing with R**. New York: Chapman & Hall, 2008.

RUE, H.; TJELMELAND, H. Fitting gaussian markov random fields to gaussian fields. **Scandinavian Journal of Statistics**: theory and applications, Stockholm, v. 29, n. 1, p. 31-49, Mar. 2002.

SCHABENBERGER, O.; GOTWAY, C. A. **Statistical methods for spatial data analysis**. New York: Chapman & Hall, 2005.

SCHLATHER, M. **Simulation and analysis of random fields**. Princeton: University, 2010. (R package version 1.3.45). Disponível em: <http://CRAN.R-project.org/package=RandomFields>. Acesso em: 18 nov. 2010.

SEARLE, S. R. **Matrix algebra useful for statistics**. New York: J. Wiley, 1982. 438 p.

SEINGH, K. On the asymptotic accuracy of efrons bootstrap.**Annals of Statistics**, Hayward, n. 9, n. 6, p. 1187-1195, Nov. 1981.

SOARES, A. **Geoestatìstica para as ciências da terra e do ambiente**. 2. ed. Lisboa: Instituto Superior Técnico, 2006. 213 p.

SOKAL, R. R.; ODEN, N. L. Spatial autocorrelation in biology 1. Methodology. **Biological Journal of the Linnean Society**, London, v. 10, n. 2, p. 199-228, June 1978a.

SOKAL, R. R.; ODEN, N. L. Spatial autocorrelation in biology 2. some biological implication and four application of evolutionary and ecological interest. **Biological Journal of the Linnean Society**, London, v. 10, n. 2, p. 229-249, June 1978b.

THOMPSON, S. K. **Sampling**. 2th ed. Oxford: Wiley-Interscience, 2002. 367 p.

TOBLER, W. R. A computer movie simulating urban growth in the Detroit region. In: International Geographcal Union (Ed.). In: **Ann Arbor**. Michigan: Commission on Quantitative Methods. August, 1969.

VENABLES,W. N.; RIPLEY, B. D. **Modern applied statistics with s**. 5th ed. New York: Springer, 2002.

WACKERNAGEL, H. **Multivariate geostatistics: an introduction with applications**. 3th ed. Berlin: Springer, 2003. 387 p.

WALLER, L. A.; GOTWAY, C. A. **Applied spatial statistics for public health data**. New Jersey: J. Wiley, 2004. 494 p.

WEBSTER, R.; OLIVER, M. A. **Geostatistics for environmental scientists**. Chichester: J. Wiley, 2001. 271 p.

WEISS, I. M. **A survey of discrete Kalman-Bucy fi ltering with unknown noise covariances**. New York: AIAA , 1970.

WHITT, W. Stochastic population model. In: KOTZ, S.; JOHNSON, N. L.; READ, C. B. (Ed.). **Encyclopedia of Statistical Sciences**. New Jersey: J. Wiley, 1986. v. 8, p. 836-851.

ZOUBIR, A. M.; BOASHASH, B. The bootstrap and its application i signal processing: An attrctive tool for assessing the accuracy of estimators and testing hypothesis for parameters in small data-sample situations. **IEEE Signal Processing Magazine**, New York, v. 15, n. 1, p. 56-76, Jan. 1998.