MINISTÉRIO DA EDUCAÇÃO

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Escola de Engenharia

Programa de Pós-Graduação em Engenharia de Minas, Metalúrgica e Materiais

PPGEM

METODOLOGIAS DE INSERÇÃO DE DADOS SOB MECANISMO DE FALTA MNAR PARA MODELAGEM DE TEORES EM DEPÓSITOS MULTIVARIADOS HETEROTÓPICOS

Camilla Zacché da Silva

Tese para obtenção do título de Doutora em Engenharia

Porto Alegre – RS

2018

MINISTÉRIO DA EDUCAÇÃO

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Escola de Engenharia

Programa de Pós-Graduação em Engenharia de Minas, Metalúrgica e Materiais

PPGEM

METODOLOGIAS DE INSERÇÃO DE DADOS SOB MECANISMO DE FALTA PARA MNAR MODELAGEM DE TEORES EM DEPÓSITOS MULTIVARIADOS HETEROTÓPICOS

Camilla Zacché da Silva

Bacharel em Física

Projeto de tese realizado no laboratório de pesquisa mineral e planejamento mineiro (LPM) da escola de engenharia da Universidade Federal do Rio Grande do Sul, dentro do programa de Pós-graduação em Engenharia de Minas, Metalúrgica e de Materiais (PPGEM), como parte dos requisitos para obtenção do título de Doutora em Engenharia.

Esta tese foi julgada adequada para obtenção de título de Doutor em Engenharia, área de concentração Metalurgia Extrativa e Tecnologia Mineral e aprovada em sua forma final pelo Orientado e Banca Examinadora do curso de Pós-Graduação

Orientador:

Prof. Dr. João Felipe Coimbra Leite Costa

Banca Examinadora:

Prof. Dra. Vanessa Koppe

Prof. Dr. Amílcar A. Soares

Dr. Diniz Ribeiro

Prof. Dr. Carlos Pérez Bergmann

Coordenador do PPGEM

“O que prevemos raramente ocorre;

o que menos esperamos geralmente acontece”

Benjamin Disraeli

À minha família

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Ao modelar-se depósitos minerais é comum enfrentarmos o problema de estimar múltiplos atributos possivelmente correlacionados, onde algumas variáveis são amostradas menos densamente do que outras. A falta de dados impõe um problema que requer atenção antes de qualquer modelagem subsequente. Precisamos, ao final, de modelos que sejam estatisticamente representativos. A maioria dos conjuntos de dados de problemas práticos são amostrados de maneira heterotópica e, para obter resultados coerentes, é preciso entender os motivos pelos quais alguns dados faltam e quais são os mecanismos que influenciaram a ausência de informações. A teoria de dados faltantes relaciona as amostras ausentes com aquelas medidas através de três mecanismos distintos: Faltante Completamente Aleatório (Missing Completely At Random - MCAR), Faltante Aleatório (Missing At Random - MAR) e Faltante Não Aleatório (Missing Not At Random - MNAR). O último mecanismo é extremamente complexo e a literatura recomenda ser tratado inicialmente como um mecanismo MAR. E após uma transformação fixa deve ser aplicada aos valores complementados para que estes se transformem em valores MNAR. Embora existam métodos estatísticos clássicos para lidar com dados faltantes, tais abordagens ignoram a correlação espacial, uma característica que ocorre naturalmente em dados geológicos. A metodologia adequada para tratar com a falta de dados geológicos é a atualização bayesiana, em que se inserem valores sob mecanismo MAR considerando a correlação espacial. No presente estudo, a atualização bayesiana foi combinada com transformações fixas para tratar o mecanismo de falta de dados MNAR em dados geológicos. A transformação fixa aqui empregada é baseada no erro de inserção gerado em um cenário MAR no conjunto de dados. Assim, com o conjunto completo resultante foi utilizado em uma simulação sequencial gaussiana dos teores de uma base de dados multivariada, apresentando resultados satisfatórios, superiores aos obtidos por meio da cossimulação sequencial gaussiana, não inserindo qualquer viés no modelo final.

Palavras-chave: MNAR, dados geológicos, simulação sequencial gaussiana, transformação fixa, cossimulação sequencial gaussiana

ABSTRACT

When modeling mineral deposits, it is common to face the problem of estimating multiple attributes possibly correlated where some variables are more densely sampled then others. Missing data imposes a problem that requires attention prior to any subsequent modeling. The later requires estimation models statistically representative. Most practical data sets are often heterotopically sampled, and to obtain coherent results one must understand the reasons why there are missing data and what are the mechanisms that cause the absence of information. The theory of missing data relates the missing samples to those measured through three different mechanisms: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). The last mechanism is quite complex to deal with, and the literature recommends being treated as a MAR mechanism and after a fixed transform should be applied to the imputed values so that these turn into MNAR imputed values. Even though there are classical statistical methods to deal with missing data, such approaches ignore spatial correlation, a feature that occurs naturally in geological data. The adequate methodology to deal with missing geologic data is Bayesian Updating, which approaches the MAR mechanism and accounts for spatial correlation. In the present study, bayesian updating was used combined with fixed transforms to treat MNAR missing data mechanism in geologic data. The fixed transform herein used is based on the error of MAR imputation on the data set. The resulting complete set was then used on a sequential gaussian simulation of the grades on a multivariate data set, presenting satisfactory results, superior to those obtained through sequential gaussian cossimulation, not inserting any biases on the final model.

Keywords: MNAR, geologic data, sequential gaussian simulation, fixed transforms, sequential gaussian cossimulation

CAPITULO 1

INTRODUÇÃO

Em avaliação de depósitos minerais é importante a quantificação, identificação e distribuição espacial de propriedades químicas e físicas do depósito. Para tanto, são realizadas campanhas de amostragens para obter esses dados geológicos. Esta coleta de dados, pode-se ater não somente a uma variável de interesse, mas múltiplas variáveis e possivelmente correlacionadas. Além disso, algumas variáveis podem estar sub amostradas em relação as outras, isto porquê muitas vezes o interesse em conhecer o comportamento de determinado atributo ocorre em etapas avançadas da campanha amostral quando por uma razão qualquer passa-se a amostrar mais elementos químicos. Portanto, ao construir um modelo da distribuição de teores do depósito, para que este seja acurado, é necessário considerar todas as relações entre os múltiplos atributos. O método geoestatístico clássico para incorporar as correlações de múltiplas variáveis é a cokrigagem (Marechal, 1970).

A cokrigagem é uma extensão da krigagem (Matheron, 1963), onde o processo de interpolação busca a minimização da variância do erro e erro médio nulo, contudo, esta é extremamente laboriosa quando se lida com mais de dois atributos, devido à necessidade de modelagem das covariâncias diretas e cruzadas entre as variáveis. Na modelagem da corregionalização multivariada de N variáveis, é necessário que sejam inferidas $\frac{N(N+1)}{2}$ covariâncias diretas e cruzadas. A dificuldade desta metodologia, porém não se restringe ao número de modelos a serem deduzidos, mas sim no fato de que os modelos não podem ser construídos independentemente uns dos outros.

Para aliviar as dificuldades operacionais com uso de cokrigagem, alguns métodos para a simplificação do uso de informações secundárias foram propostos, dentre estes: krigagem com médias locais, krigagem com tendência externa e Modelos de Markov (Almeida, 1993; Journel, 1999). Todavia, essas metodologias apresentam uma limitação, necessitam que a informação secundária seja exaustivamente amostrada ou que seja feita sua complementação por meio de inserção de dados (Isaaks e Srivastava, 1989).

Outra maneira de analisar conjuntos de dados multivariados em que as variáveis apresentam algum tipo de correlação entre si, seja espacial ou temporal, são os métodos de fatoração, que consistem em decomposições espectrais dos vetores de dados, a fim de descorrelacioná-los e, desta forma, estudar cada fator separadamente. Neste contexto, tem-se o método de Análise das Componentes Principais (*Principal Component Analysis – PCA*) (Pearson, 1901; Hoettling,1933), que consiste em realizar uma rotação no sistema de coordenadas das observações correlacionadas, levando-as a um novo sistema de referências em que as variáveis são ortogonais entre si (correlação igual à zero). Procede-se ao processamento de cada componente de maneira individual utilizando geoestatística univariada, e, depois de modeladas, a retrotransformação as traz ao sistema de referência original sem perda de informação (Davis e Greenes, 1983). Contudo, a descorrelação acontece somente para vetores de separação nulos, ou seja, para amostras colocadas de variáveis distintas (Davis e Greenes, 1983).

Outro método conhecido é o de Fatores de Autocorrelações Mínimas e Máximas (*Minimum-Maximum Autocorrelation Factors* – MAF*)* (Switzer e Green, 1984). Este método é derivado do PCA, e, como tal, também leva as variáveis originais a fatores descorrelacionados, entretanto, o MAF obtém a descorrelação das variáveis para vetores de separação não nulos. Desta forma, é possível trabalhar cada variável do sistema de maneira univariada (descorrelacionada).

Os métodos de fatoração citados, apesar de apresentarem resultados promissores (Silva, 2013) ao evitar o uso de modelos de corregionalização impostos pela metodologia clássica, há ainda a necessidade de que todas as variáveis em análise sejam igualmente amostradas em todos os locais, em outras palavras, os métodos de fatoração exigem isotopia do banco de dados.

Metodologias de fatoração tem sua aplicação cada vez mais frequente, porém abordagens como PCA e MAF garantem a descorrelação, mas não a independência das variáveis, ou seja, falham ao lidar com relações multivariadas complexas. Desta forma, transformações multigaussianas como Stepwise Conditional Transform (SCT) (Rosemblat, 1952; Leuangthong e Deutsch, 2003) e Projection Persuit Multivariate Transform (PPMT) (Barnett et. al, 2014) surgem a fim de atender algumas das limitações das metodologias de PCA e MAF. A última surge com a proposta de, além de considerar as relações multivariadas complexas, lidar com conjuntos de dados com grande número de variáveis e amostras.

A transformação PPMT tem por objetivo mapear uma distribuição multivariada não gaussiana em uma distribuição de igual dimensão gaussiana e independente, de modo que o uso do modelo linear de corregionalização pode ser evitado. Contudo, apesar de ser um avanço em

relação às abordagens supracitadas, essa também necessita de amostragem isotópica (Barnett e Deutsch, 2012).

Na prática, bancos de dados isotópicos são raros e por isso o usuário recai sobre duas alternativas: descartar amostras incompletas ou realizar a complementação (inserção) das faltantes. A primeira leva à diminuição da quantidade de informação utilizada para a construção do modelo de teores, e pode levar a resultados que possuem viés estatístico. A segunda deve ser aplicada de forma que respeite a variabilidade dos dados, a estatística local condicional e sua correlação espacial para que ao fim, o modelo construído reproduza o fenômeno observado. Assim, a busca por metodologias que realizem a inclusão de dados de maneira sã torna-se pertinente.

1.1 Dados faltantes

A incompletude de informação em uma base de dados é um problema recorrente em diversas áreas de pesquisa como médica, social, além da área de mineração. Por décadas, vem-se aplicando técnicas para solucionar os problemas gerados pela ausência de dados incompletos, por exemplo, exclusão de amostras incompletas ou inserção de valores às amostras incompletas, que se baseiam em hipóteses bastante restritas e potencialmente adicionarão viés às estimativas e análise final do conjunto de dados.

Assim, heterotopia amostral tem sido objeto de estudo há mais de um século (Enders, 2010); porém, foi em 1976, quando Rubin esboçou a teoria de inserção de dados e classificou as relações entre dados faltantes e variáveis medidas que houve os grandes avanços na matéria, permitindo entender o porquê algumas técnicas se mostravam adequadas a determinados casos e outras pareciam falhar quando aplicadas.

Existem três mecanismos principais que descrevem a natureza da falta de dados de acordo com Rubin (1976), **faltantes aleatórios** (*Missing at Random – MAR*), **faltantes completamente aleatórios** (*Missing Completely at Random – MCAR*) e **faltantes não aleatórios** (*Missing Not at Random – MNAR*). A cada caso de incompletude de dados há uma maneira de tratá-la, de modo que a inserção de dados não gere viés significativo às estimativas finais.

1.1.1 Faltantes aleatórios (Missing at Random - MAR)

Dado um vetor de observações $\Phi[Z_1, Z_2, ..., Z_N]$ com N variáveis, é considerado o mecanismo MAR, se a probabilidade de faltar um valor da variável Z_1, por exemplo, depende de alguma das outras N -1 variáveis do modelo e não dos valores de Z_1 propriamente ditos. É importante ressaltar que apesar de sua nomenclatura, essa hipótese não se assemelha a um

comportamento randômico, como o de um jogo de cara ou coroa, mas significa que há uma relação sistemática entre as variáveis medidas e a probabilidade de Z_1 ser incompleta na amostra. Por exemplo, caso seja realizada uma campanha amostral em um depósito mineral e dentre as variáveis coletadas duas, Z_1 e Z_2 apresentem forte correlação, Z_1 deixa de ser amostrada e posteriormente será estimada via regressão a fim de diminuir os custos de amostragem. Onde há teores da variável Z_2, a probabilidade que Z_1 não seja amostrada é maior. Já no caso em que Z_1 não possui amostras em todos locais por outro motivo que não os teores de Z_2, por exemplo, a variável se concentra em maior parte em locais de solo friável e, portanto, não é possível a recuperação do testemunho, a hipótese de MAR fica sem a possibilidade de verificação, sendo este um problema prático deste mecanismo, pois algumas técnicas de inserção de dados como **máxima verossimilhança** (*Maximum likelyhood*) assumem esta hipótese.

1.1.2 Faltantes completamente aleatórios (Missing Completely at Random - MCAR)

A definição do mecanismo de faltantes completamente aleatórios (MCAR) é que a probabilidade de faltar um valor da variável Z_1 não depende dos seus próprios valores e de nenhuma das outras N-1 variáveis do modelo. Isto torna essa hipótese ainda mais restritiva que o MAR, pois ela assume que não existe nenhum tipo de relação entre dados e falta de dados (Enders, 2010). Por exemplo, dados faltantes sob o mecanismo completamente aleatório podem ser vistos como aqueles em que a não amostragem de um local não depende da intenção da coleta de dados ou mesmo de alguma característica geológica, mas como decorrente de decisões tomadas ao longo do tempo, como a expansão da área de mineração, ou ainda inacessibilidade do local devido ao bioma da região.

A definição de MCAR implica que os casos de dados faltantes pertencem à mesma população, ou seja, compartilhem o vetor de médias e a matriz de covariâncias, já que o mecanismo exige que os dados observados sejam uma simples amostra aleatória de um conjunto hipoteticamente completo. Kim e Bentler (2002) se referem a este comportamento como homogeneidade de médias e covariâncias.

A vantagem deste mecanismo, apesar de mais restritivo, é a possibilidade de testar esta chamada homogeneidade. Dixon (1988) propõe um teste do MCAR, chamado *t-test* e Little (1988) propõe uma extensão multivariada do teste proposto por Dixon (1988).

1.1.3 Faltantes não aleatórios (Missing Not at Random - MNAR)

Quando a probabilidade de faltar um dado da variável Z_1 depende dos valores de Z_1 e de alguma das outras N-1 variáveis do conjunto em análise. Por exemplo, considere que em um depósito de ouro há indícios que em determinada área há ocorrência somente de baixos teores de ouro. Assim, por motivos econômicos não são coletadas amostras de ouro. Desta maneira, a ausência de informação decorre do próprio valor de ouro. Ou ainda, que em uma pesquisa médica sobre a qualidade de vida de pacientes em tratamento de câncer. Suponha, que alguns pacientes durante a pesquisa se tornem tão doentes, que são impossibilitados de continuar participando do estudo. Neste caso; portanto, a falta do dado sobre a qualidade de vida do paciente depende diretamente da atual qualidade de vida do paciente (Enders, 2010). Como no caso de faltantes aleatórios, esta hipótese não é passível de testes, uma vez que o teste depende dos valores faltantes (Enders, 2010).

Quando Rubin esboçou a sua teoria de dados faltantes explicitou as condições que devem ser atendidas para que os parâmetros estatísticos da população sejam estimados adequadamente. A teoria de inserção de dados mostra que as metodologias utilizadas historicamente, como análise de caso completo, podem apenas ser aplicadas à conjuntos em que o mecanismo de falta é MCAR sem a adição de viés às estimativas finais dos parâmetros. Porém, esse mecanismo não rege a maioria dos casos práticos de modo que, o mecanismo que rege a ausência de dados deve ser propriamente identificado a fim de que não sejam conferidos erros às previsões de variáveis. Portanto, é necessário que a abordagem aplicada para inserção de dados ao conjunto deve realizada de acordo com o mecanismo identificado. A seção 1.2 apresenta as metodologias clássicas utilizadas para inclusão de valores no conjunto amostral adequada a cada mecanismo de falta de dados.

1.2 Metodologias aplicada aos mecanismos de falta MCAR e MAR.

1.2.1 Metodologias clássicas aplicadas a bancos de dados incompletos

Em seus primórdios, o estudo sobre dados faltantes propôs diversas maneiras de lidar com bancos de dados incompletos baseando-se em descarte de amostras incompletas. Estas metodologias encontraram grande aceitação devido à sua rápida aplicabilidade aos dados. Porém, este tipo de abordagem mostrou-se ineficiente (Little e Rubin, 2002) ao verificar que viés é inserido aos resultados e que somente a casos onde o mecanismo de falta regente é MCAR não se obtém viés nas estimativas. Neste âmbito, recaem a análise de caso completo (listwise

deletion) e a análise de pares (pairwise). A análise de caso completo**,** consiste em descartar todo e qualquer caso que tenha um ou mais dados faltantes. O benefício desta técnica é a conveniência, pois se restringe a análise de casos completos, porém a restrição de intervalo dos dados irá diminuir a variabilidade da distribuição dos dados, o que reduzirá a magnitude da correlação existente entre variáveis. De acordo com Watanabe et al. (2009), a medida que o coeficiente de correlação entre as variáveis diminui, os processos de coestimativa tornam-se dominados pela variável secundária, sendo esta mais abundante, podendo acarretar em viés nas estimativas da variável primária.

Há também, o método de **inserção de dados pela média não condicional** ou **inserção da média aritmética** que aplica a substituição de valores faltantes pela média aritmética dos dados disponíveis (Wilks, 1932). Este método é amplamente usado, pois produz rapidamente um banco de dados completos. Apesar disso, esta metodologia produz fortes distorções nas estimativas dos parâmetros da distribuição dos dados, se o mecanismo que governa a falta não é MCAR. Além disso, fica claro que a inclusão da média em locais de informação faltante atenua a variabilidade da distribuição, afetando também as correlações e covariâncias entre variáveis.

A **inserção por regressão linear** ou **inserção condicional da média**, como também é conhecida, é aplicada a casos onde as variáveis apresentam boa correlação, usando a informação completa para gerar valores às amostras incompletas. A ideia básica deste método é estimar uma equação de regressão ou um conjunto de equações e a partir destas prever os valores faltantes do banco de dados. Em 2011, Reuwsaat realizou um estudo de previsibilidade de teores em um banco de dados constituído por 6 variáveis principais, Fe SiO_2, Al_2O_3, Mn, P e PF, o qual apresenta heteropia parcial entre as seis variáveis, ou seja, há amostras incompletas no conjunto de dados. Neste estudo, Reuwsaat (2011) apresenta a regressão linear como alternativa de isotopização, como chama em seu texto, para o banco de dados previamente à estimativa por meio de cokrigagem do conjunto de amostras. Ele mostra neste estudo que a utilização da informação abundante obtida por meio de regressão linear mostra-se superior a exclusão de amostras incompletas, onde o modelo de blocos obtidos ao final do procedimento é validado de acordo com um modelo de referência calculado previamente, com distribuição de erro bastante similar ao modelo de referência. Este método é bastante rápido; porém, quando há múltiplas variáveis com dados faltantes a utilização da regressão linear pode se tornar complexa.

Inserção por regressão estocástica assim como a inclusão condicional da média, utiliza um conjunto de equações de regressão para prever os valores desconhecidos do banco

de dados a partir das amostras completas. A complementação da informação faltante por meio de regressão estocástica, porém vai além no sentido de que recupera a variabilidade dos dados perdida na estimativa por regressão adicionado um termo residual que possui distribuição normal.

1.2.2 Estimativa por máxima verossimilhança

Estimativa por máxima verossimilhança (Maximum Likelihood Estimation – MLE) é o método tomado como estado da arte em estatística moderna para análise de dados faltantes. O ponto de partida do método é especificar a distribuição que melhor ajusta a população de amostras. Caso os dados possuam distribuição gaussiana, esta é completamente definida pela média e variância. O objetivo do método torna-se, portanto, a estimativa dos parâmetros da distribuição dos dados.

Por exemplo, no caso univariado onde a variável Z possui z_i amostras, a função densidade de probabilidade é dada pela equação (1):

$$L_i = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} e^{\frac{-0.5(z_i-\mu)^2}{\sigma^2}} \qquad (1)$$

onde z_i é um valor amostral de Z, μ é a média e σ^2 a variância da distribuição de Z, L_i é a probabilidade relativa de se obter este valor amostral z_i a partir da distribuição dos dados. Desta forma, se busca o par, variância e média, para a distribuição que produza a maior probabilidade relativa. Uma vez determinados os parâmetros da distribuição que melhor ajusta os dados, as amostras faltantes podem ser preenchidas por meio de simulação estocástica.

Na aplicação da metodologia a bancos de dados incompletos, a função de densidade de probabilidade exige pequenas alterações a fim de acomodar o fato de que as múltiplas variáveis não possuem o mesmo número de amostras. Há algoritmos que buscam a otimização dos parâmetros de distribuição considerando as nuances presentes em bancos de dados incompletos, tais como o algoritmo de Maximização de Expectativa (Expectation Maximization – EM).

A metodologia MLE pode ser indicada para aplicação em situações que o mecanismo que domina a falta de dados é MAR, pois realiza estimativas dos parâmetros sem viés significativo quando esta hipótese é satisfeita. A sua aplicação, quando o mecanismo de falta é

MCAR, hipótese exigida pelas metodologias apresentadas anteriormente a esta, mostra-se superior em suas estimativas.

1.2.3 Inserção múltipla

Assim como MLE, a inserção múltipla (Multiple Imputation – MI) assume que os dados são normalmente distribuídos e que o mecanismo que rege a falta dos dados é MAR. Porém, de forma diferente do MLE, a múltipla inserção cria diversos bancos de dados completos e únicos, enquanto MLE estima de fato a distribuição de um banco de dados hipoteticamente completo. MI é dividida em três fases: fase de inserção, fase de análise e fase de combinação. De modo que, a primeira fase é fortemente calcada na abordagem bayesiana desenvolvida por Rubin (1987), que será explicada em detalhes no Capitulo 2.

A fase de inserção de MI é dividida em duas etapas: passo I e passo P.

- Passo I: abordagem próxima à regressão estocástica utilizada para construir um conjunto de equações de regressão para formar um banco de dados completo, com os faltantes estimados a partir dos presentes.
- Passo P: o objetivo deste passo é gerar estimativas do vetor de médias e matrizes de covariâncias para o banco completo formado no passo I. De modo que, as estatísticas atualizadas irão alimentar o passo I.

A partir destes dois passos, a fase de inserção se completa gerando *p* bancos de dados completos e únicos, formados por estatísticas diversas, geradas no passo P. Após a realização dos passos I e P, vem a fase de análise, que, de forma simples, realiza a análise estatística de cada cenário gerado na fase de inserção. A fase de combinação; porém, é mais complexa, pois o objetivo desta é realizar um único cenário resultante a partir dos diversos gerados inicialmente. Rubin (1987) definiu como combinação das diversas inserções realizadas a média aritmética das mesmas. Em relação aos desvios padrões ou variâncias, a combinação se divide em duas componentes, pois a inserção múltipla considera duas fontes de flutuação amostral. A variância interna do cenário complementado (within imputation variance) e a variância entre cenários complementados (among imputation variance). A variância interna do cenário complementado descreve a variância amostral caso o banco de dados de entrada fosse completo. De forma intuitiva, os dados faltantes originalmente deveriam aumentar a variância dos dados. Assim, quando é considerado somente um cenário complementado (inserção única) o desvio padrão é subestimado, pois os valores atribuídos são tomados como reais. A inserção múltipla

considera mais uma fonte de variabilidade amostral, a variância entre cenários complementados, restaurando as variâncias dos dados caso o banco fosse completo. Assim como MLE, este método propõe a inserção de dados sem viés em conjuntos em que o mecanismo de falta atuante é MAR. Ainda, a metodologia propõe a restauração da variabilidade dos dados caso estes formassem um conjunto completo, característica que não é abordado na metodologia do item 1.2.2.

1.2.4 Inserção por atualização bayesiana

As metodologias até agora apresentadas não levam em consideração correlação espacial entre os dados, característica presente em dados geológicos. Portanto, faz-se necessária uma metodologia que considere este tipo de relação entre as amostras, a fim de que a informação adicional, obtida através de inserção de dados faltantes, utilizada no modelo para estimativa de teores seja a mais acurada possível. Ren (2007) propõe uma abordagem baseada em MI, porém considerando a correlação espacial entre a variável com amostras faltantes e uma série de variáveis secundárias completas, colocadas e correlacionadas entre si.

O método por ele proposto utiliza estimativa bayesiana, assim como a primeira fase da inserção múltipla. O método consiste em construir distribuições condicionais nos locais onde é necessária a inclusão de valores, assumindo que o conjunto é multivariado gaussiano. A variável incompleta é tratada como primária e as restantes como secundárias. O procedimento de estimativa bayesiana utiliza três distribuições: primária, neste caso obtida a partir das equações de krigagem simples da variável que possui dados faltantes; a distribuição de probabilidade, que por sua vez é resultante da coestimativa utilizando os atributos secundários; e a posterior, esta última é uma combinação das duas primeiras. Uma vez que a distribuição posterior é definida, os valores faltantes podem ser obtidos por meio de amostragem aleatória desta, utilizando simulação estocástica. Este método mostrou ganho de precisão quando comparado com a abordagem da análise de caso completo. Além disso, a metodologia torna-se especialmente atraente devido ao tratamento que esta dá a dados espacialmente correlacionados. A limitação desta abordagem, porém é considerar que os dados são multivariados gaussianos, e tratando-se de dados geológicos, em sua maioria, apresentam características estatísticas mais complexas, onde a multigaussianidade não pode ser assumida.

1.3 Metodologias aplicadas ao mecanismo de falta MNAR

Quando o mecanismo de falta regente dos dados é MNAR as metodologias supracitadas não podem ser utilizadas, uma vez que assumem a hipótese de MAR. Assim, produzem estimativas que carregam viés quando MAR não caracteriza a probabilidade de falta de dados.

Rubin (1976) propõe em sua teoria de dados faltantes, que a probabilidade de falta de amostras seja analisada a partir de uma variável indicadora, R, de falta, que descreve se o dado é faltante (R=0) ou presente (R=1).

No mecanismo de falta MNAR, os parâmetros da distrobuição de dados faltantes carregam informações importantes de modo que não podem ser ignorados. As metodologias indicadas para abordar conjuntos com tal característica procuram considerar a importância destas informações, incorporando um modelo que descreva a relação dos dados medidos e a propensão de falta de dados. Há duas metodologias calcadas na distribuição conjunta dos dados e probabilidade de falta de dados: Modelo de Seleção e Modelo de Mistura de Padrões.

O modelo de seleção define a distribuição conjunta de dados e probabilidade de falta de dados como o produto de duas distribuições componentes:

$$p(Z,R) = p(R|Z)\,p(Z) \qquad (2)$$

onde $p(R|Z)$ é a distribuição condicional de falta dado Z, e $p(Z)$ é a distribuição marginal dos dados. A equação 2 descreve a probabilidade de se obter diferentes valores de Z bem como determina a probabilidade da amostra estar incompleta.

O modelo de mistura de padrões propõe um produto distinto de duas distribuições componentes:

$$p(Z,R) = p(Z|R)\,p(R) \qquad (3)$$

Onde $p(Z|R)$ é a distribuição condicional dos dados para determinado R e $p(R)$ a distribuição marginal da falta de dados. A distribuição conjunta neste caso descreve a probabilidade de se obter diferentes valores de Z dentro de um subconjunto de amostras incompletas que apresentam o mesmo padrão de falta.

Os dois modelos apresentados são intercambiáveis, pois procuram descrever o mesmo fenômeno. Porém, as duas abordagens dependem da modelagem da distribuição conjunta *P(Z,R)* e estimar tal modelo é uma tarefa árdua, uma vez que nos casos práticos não se sabe porquê os dados são faltantes. Além disso, o mecanismo de falta MNAR não é uma hipótese passível de teste.

Contudo, Rubin (1987) propõe uma análise de sensibilidade posterior aos processos de inserção aplicáveis à hipótese MAR. A aplicação proposta por Rubin (1987) é simples: existindo razões para crer que o mecanismo de falta atuante no conjunto amostral seja MNAR, criam-se valores complementados por meio de métodos desenvolvidos para mecanismo MAR. Após, aplicar-se-á uma transformação fixa aos valores complementados a fim de transformá-los em saídas MNAR. As análises estatísticas e geoestatísticas devem ser realizadas após a aplicação da transformação fixa, examinando o modelo gerado frente ao conhecimento do comportamento dos atributos. Assim, torna-se interessante aplicar diferentes transformações fixas ao conjunto complementado, para que diferentes modelos sejam gerados e o analista possa julgar qual é o modelo mais adequado.

Em 1987, Rubin propõe uma série de transformações fixas às saídas dos modelos MAR, dentre elas, a adição de uma constante de 20% de cada valor complementado. Nesta linha, Cohen (1988) sugere que a constante seja metade do valor do desvio padrão do conjunto complementado, tornando essa constante uma métrica de efeito sobre a média do conjunto. Tal abordagem torna-se atraente devido à praticidade de aplicação e fácil comunicação entre diversos modelos gerados.

No contexto geológico, ainda é um desafio o caso de bancos de dados heterotópicos, visto que há inúmeros casos em que os dados faltantes são MNAR. Desta maneira, pesquisas devem ser consideradas neste âmbito a fim de que os modelos gerados sejam representativos do fenômeno em estudo, uma vez que o mecanismo MNAR afeta fortemente a distribuição da amostra de dados.

1.4 Problema

Em depósitos minerais multivariados, é necessário determinar a quantidade e qualidade do minério e os teores de seus múltiplos elementos. O processo de modelagem geoestatística multivariada pode ser extremamente laborioso com o número crescente de variáveis, de modo que por muitas vezes são utilizadas metodologias para remover as correlações entre os atributos de estudo, tratando-os individualmente.

Tais metodologias, como PCA, MAF e PPMT por sua vez, exigem que todas as variáveis estejam amostradas nos mesmos locais (amostragem isotópica). Atualmente, se recorre à eliminação de amostras incompletas e trabalha-se com um subconjunto dos dados completos para utilizar-se tais métodos. A limitação desta abordagem é evidente, uma vez que é inevitável a perda de informação ao longo da modelagem. Além disso, em dados geológicos

há casos de conjuntos incompletos que se encontram sob o mecanismo de falta MNAR, o que leva às distribuições que não são representativas do atributo em análise. Logo, ainda que o estudo não busque a descorrelação dos atributos, e a coestimativa (ou cossimulação) seja utilizada, ainda há a necessidade de se dispor distribuições representativas do conjunto de dados. Portanto, as metodologias propostas para análise geoestatística de teores em depósitos multivariados podem ser usadas adequadamente somente quando combinadas a um processo de inserção de dados, que não leve a perda de informação e não gere viés no conjunto amostral em estudo.

1.5 Meta

O presente trabalho propõe explorar e desenvolver procedimentos de inserção de valores em conjuntos que estejam submetidos ao mecanismo de falta MNAR, eliminando a heterotopia dos dados sem que viés seja adicionado ao modelo. Ao fim deseja-se utilizar o conjunto completo gerado em análises geoestatísticas univariadas ou de fatoração para remoção de correlação entre atributos.

Declaração de tese:

A metodologia ideal busca por meio da inserção de valores gerar um conjunto completo que respeite a correlação espacial apresentada pelo fenômeno geológico, que seja de rápida implementação computacional e que o modelo resultante seja representativo quando o mecanismo de falta for MNAR.

As condições apresentadas não são satisfeitas simultaneamente pelas metodologias até o momento aplicadas ao problema de dados geológicos faltantes. Portanto, a meta deste trabalho é obter uma metodologia que trate o mecanismo de falta MNAR em dados geológicos e implementar o procedimento proposto, e por meio de análise exploratória verificar a aplicabilidade deste a casos multivariados complexos.

1.6 Contribuição da tese

A contribuição da tese se dá pela busca de uma metodologia, utilizadas em outras áreas do conhecimento, e adaptadas ao problema de modelagem de teores em depósitos polimetálicos complexos que formam um conjunto amostral heterotópico.

Desta maneira, procedimentos embasados na cokrigagem, que aceita banco de dados heterópicos, porém demanda a utilização do modelo linear de corregionalização e modelos univariados aplicados à conjuntos complementados por meio de um processo de inserção serão desenvolvidos e comparados a um conjunto de dados exaustivo a fim de avaliar a performance dos modelos gerados.

1.7 Objetivos

Para atingir a meta proposta são considerados os seguintes objetivos específicos:

I. Investigar metodologias que considerem o mecanismo de falta de dados não aleatória (MNAR);
II. Desenvolver (adaptar) uma metodologia que considere o mecanismo MNAR;
III. Realizar a análise comparativa dos resultados obtidos por meio das diversas metodologias (métricas), observando a coerência e reprodução das características originais das variáveis em estudo;
IV. Analisar comparativamente os modelos simulados por meio de simulação sequencial gaussiana dos conjuntos de dados complementados bem como gerar um modelo cossimulado do conjunto de dados hetorotópico;
V. A partir do conjunto exaustivo de referência, avaliar o desempenho da metodologia proposta a fim de determinar se a solução é apropriada no que se refere à modelagem de teores, bem como vantagens e limitações de cada um dos procedimentos expostos.

1.8 Metodologia

Primeiramente, será gerado intencionalmente um banco de dados incompleto sob mecanismo MNAR, em um cenário controlado onde haverá o referencial completo para fins de validação estatística. É necessário que o banco de dados utilizado seja multivariado.

A primeira etapa do estudo consiste em realizar o procedimento de inserção de dados no conjunto incompleto gerado. A inserção será realizada por meio de atualização bayesiana aplicada a dados geológicos. Isso constitui um procedimento de inserção múltipla considerando a correlação espacial entre as variáveis por intermédio das distribuições primária, probabilidade e posterior. Este procedimento gera o banco de dados completo assumindo o mecanismo MAR.

Após os conjuntos de dados completos previamente gerados serão submetidos às transformações fixas (Rubin, 1987), tratando desta forma o mecanismo de falta MNAR. Assim,

o fim da primeira etapa resulta em diversos modelos, o obtido por meio de atualização bayesiana sem qualquer transformação fixa aplicada, e os modelos obtidos após a aplicação das transformações fixas. Posteriormente, estes modelos serão confrontados ao conjunto referência, a fim de avaliar o desempenho de cada modelo e compreender as condições em que a aplicação das transformações fixas é adequada.

A segunda etapa desta tese aborda a construção de um modelo de incerteza dos teores de um conjunto de dados multivariado por meio de Simulação Sequencial Gaussiana (SSG) (Isaaks, 1990). A SSG assume a gaussianidade dos dados, e a partir disso, sorteia aleatoriamente valores condicionados à função distribuição cumulativa condicional (*conditional cumulative distribution function - ccdf*) gaussiana, definida pela média e variância obtidas por meio de krigagem simples (Matheron, 1962). Cada valor simulado corresponde a um sorteio de valor da ccdf definida no local. A SSG será aplicada aos conjuntos de dados complementados submetidos às transformações fixas (Rubin, 1987).

Além da SSG, será aplicada a cossimulação sequencial gaussiana (COSSG) (Verly, 1993). Por meio da COSGS, é possível utilizar o conjunto de dados heterotópico para a construção do modelo geoestatístico de teores. Neste caso, o processo assume multi gaussianidade dos dados e a ccdf é construída a partir da média e variância da cokrigagem simples. Considerar atributos heterotópicos correlacionados no processo de simulação leva à construção de modelos de covariâncias que devem satisfazer à condição de definição positiva do Modelo Linear de Corregionalização (Goovaerts, 1997).

A terceira etapa desta tese consiste na avaliação da aplicabilidade e desempenho dos modelos construídos na segunda etapa.

1.9 Organização da tese

A tese está subdividida nos seguintes capítulos e desta forma organizada: o capítulo 2 apresenta uma breve revisão bibliográfica da análise de dados faltantes e visita as metodologias estatísticas clássicas para abordagem do tema. Apresenta também a adaptação realizada por (Doyen et al., 1996; Ren, 2007) à inserção múltipla considerando a correlação espacial, característica de dados geológicos.

O capítulo 3 engloba a primeira etapa da metodologia apresentada, onde o banco de dados utilizado, Jura (Goovaerts, 1997) é exposto por meio de análise estatística e descrição geológica do mesmo. Este capítulo compreende também metodologia desenvolvida para tratar

dados faltantes sob mecanismo MNAR. É exposto neste mesmo capítulo a análise e discussão dos resultados para o estudo de caso realizado.

O capítulo 4 mostra o estudo de caso realizado no conjunto Walker lake (Isaaks e Srivastava, 1989), em que foram gerados os modelos geoestatístico de teores por meio de SGS combinada a inserção de dados para casos MNAR e o modelo COSGS aplicado ao conjunto de dados heterotópico.

O capítulo 5 apresenta as conclusões obtidas ao longo do desenvolvimento desta tese além das recomendações para trabalhos futuros.

CAPITULO 2

Fundamentos teóricos: - análise de dados faltantes e inserção de dados

Este capítulo trata de uma revisão teórica dos métodos clássicos estatísticos de inserção de dados bem como dos métodos atualmente desenvolvidos na área.

2.1 Estimativas por Máxima Verossimilhança (Maximum Likelihood Estimation)

Em geoestatística se procura conhecer o comportamento e as leis que governam os fenômenos geológicos em estudo, de modo que possa ser predito o valor de um local não amostrado a fim de obter um modelo dos recursos presentes, permitindo desta forma que as diversas áreas que abrangem um empreendimento mineiro possam ser planejadas adequadamente a realidade do depósito mineral.

O ponto de partida para estimativa por máxima verossimilhança é especificar a distribuição para a população de dados (Enders, 2010), atualmente as aplicações da metodologia assumem que o conjunto de dados possui distribuições normais, de modo que são completamente especificadas pela sua média e variância.

A metodologia é calcada na distribuição de probabilidade do conjunto de dados, esta distribuição é descrita pela função densidade de probabilidade, que para uma variável gaussiana é dada por:

$$G_i = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-0.5(z_i-\mu)^2}{\sigma^2}} \qquad (4)$$

Onde z_i é um valor amostral e μ é a média da população, σ^2 é variância da população, G_i é a probabilidade, esta função descreve a probabilidade relativa de se obter um determinado valor amostral a partir da distribuição normal dos dados, sendo este o objetivo da estimativa por máxima verossimilhança.

Na prática se está interessado não na probabilidade de se obter um valor amostral da distribuição, mas sim um conjunto de valores amostrais, em teoria da probabilidade isto se obtém por meio do produto das probabilidades individuais, ou seja, a probabilidade conjunta de eventos independentes ocorrerem.

$$G = \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-0.5(z_i-\mu)^2}{\sigma^2}} \right\} \qquad (5)$$

A equação 5 quantifica a probabilidade conjunta de se obter uma coleção de N valores conhecidos a partir da mesma distribuição normal. Os valores resultantes deste produtório são de pequena magnitude, de modo que se torna mais fácil e menos suscetível a erros trabalhar com o logaritmo natural deste valor, mantendo significado. Uma regra de logaritmos, por exemplo, é que o logartimo do produto é a soma dos logaritmos, ou seja:

$$\log(XY) = \log(X) + \log(Y) \qquad (6)$$

Deste modo, o produtório da equação 5 se torna um somatório.

$$\log G = \sum_{i=1}^{N} \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-0.5(z_i-\mu)^2}{\sigma^2}} \right\} \qquad (7)$$

Assim, quando os parâmetros da população são desconhecidos é possível estima-los a partir de um processo iterativo que "testa" diversos valores de média e variância a fim de encontrar o par que produza o melhor ajuste a população de dados, ou seja, que maximize o log da probabilidade conjunta dos dados.

Testar muitos valores de médias e variâncias pode-se tornar um processo demasiadamente laborioso, assim se faz uso de derivadas parciais. A fim de encontrar o máximo de uma função, se calcula a derivada primeira desta função e igualando a equação à zero pode-se isolar o parâmetro desconhecido e obter o valor em que esta função atinge seu pico.

A aplicação prática do método de estimativa por máxima verossimilhança é a conjuntos de dados multivariados, e sua extensão é bastante direta. Analogamente, se busca a distribuição normal multivariada que melhor ajuste o conjunto amostral. Tomando $\boldsymbol{\mu}$ como o vetor de médias e $\boldsymbol{\Sigma}$ a matriz de covariâncias das variáveis e $\boldsymbol{\Phi}_\mathbf{i}$ o vetor randômico de observações, a função de densidade de probabilidade multivariada é:

$$G_i = \frac{1}{(2\pi)^{n/2} \left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-0.5(\boldsymbol{\Phi}_\mathbf{i}-\boldsymbol{\mu})^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\Phi}_\mathbf{i}-\boldsymbol{\mu})} \qquad (8)$$

onde n é o número de amostras pertencentes a cada variável. Assim como para o caso univariado, o logaritmo da função densidade de probabilidade é menos suscetível a erros de cálculo, portanto a generalização multivariada:

$$\log G_i = \log\left\{ \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-0.5(\boldsymbol{\Phi}_i-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Phi}_i-\boldsymbol{\mu})} \right\}$$
$$= \log G_i = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{\Phi}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Phi}_i - \boldsymbol{\mu}) \quad (9)$$

Sendo esta a equação que representa o logaritmo das funções de densidade de probabilidade multivariada. De maneira consistente com o caso univariado, o logaritmo da função densidade de probabilidade multivariada do conjunto amostral é o somatório das funções densidade de probabilidade, ou seja,

$$\log G = \sum_{i=1}^{N} \log G_i \quad (10)$$

Uma vez determinados os conceitos básicos do método de estimativa por máxima verossimilhança, este pode ser aplicado à análise de dados faltantes. Como foi visto a metodologia é utilizada para estimar os parâmetros de uma distribuição que melhor ajuste a população de dados, e o procedimento em essência é o mesmo com um conjunto de dados completos ou incompletos. Porém quando o conjunto de dados é incompleto, ou heterotópico, algumas alterações devem ser realizadas na função densidade de probabilidade a fim de considerar o fato de as variáveis não terem o mesmo número de amostras.

Quando o mecanismo de falta de dados é MAR, a estimativa por máxima verossimilhança é considerado o estado da arte (Schafer e Graham, 2002; Enders, 2010), pois esta produz estimativas dos parâmetros de distribuição livres de viés. Para casos de dados incompletos considere a equação 9, onde n agora irá representar o número de amostras completas do banco de dados. Isto é importante, pois neste caso é descrita a possibilidade de variação de tamanho e valores entre indivíduos, de tal forma que o log da função densidade de probabilidade depende somente das amostras completas. Assim, estimar os parâmetros da população como o vetor de médias e a matriz de covariâncias, exige um processo iterativo, que quando atinge a sua convergência produz os resultados que maximizam a função densidade de probabilidade (Enders, 2010).

O Algoritmo EM é um procedimento que realiza esta iteração necessária para a estimativa de parâmetros em conjuntos amostrais incompletos, que auxilia na análise de dados faltantes, bem como pode ser utilizado como uma ferramenta para inserção de dados.

O algoritmo EM (Dempster et al., 1977), é um processo iterativo em dois passos: o passo E e o passo M (esperança e maximização). O passo E consiste em construir um conjunto de

equações de regressão estocástica a partir dos elementos do vetor de médias e da matriz de covariância. A partir das equações de regressão prevê as amostras incompletas a partir das completas, assim preenchendo os dados faltantes. O passo M então utiliza o conjunto preenchido anteriormente no passo E e realiza uma nova estimativa do vetor de médias e da matriz de covariâncias. O algoritmo leva os novos valores do vetor de médias e da matriz de covariância para o próximo passo E que novamente calcula um conjunto de regressões estocásticas e novamente preenche os dados faltantes, em seguida o passo M realiza uma nova estimativa dos parâmetros da distribuição. Estes dois passos se repetem até que seja atingida a convergência, obtendo o vetor de médias e a matriz de covariância que maximizam a função de densidade de probabilidade multivariada (Enders, 2010).

Caso o banco de dados seja completo o conjunto de formulas a seguir levam valores de média, variância e covariância:

$$\begin{cases} \mu_{Z_i} = \dfrac{1}{n}\sum Z_i \\ \sigma^2_{Z_i} = \dfrac{1}{n}\left(\sum Z_i^2 - \dfrac{\left(\sum Z_i\right)^2}{n}\right) \\ Cov_{Z_iZ_j} = \dfrac{1}{n}\left(\sum Z_i Z_j - \dfrac{\sum Z_i \sum Z_j}{n}\right) \end{cases} \qquad (11)$$

Onde μ_{Z_i} representa a média da variável Z_i e $\sigma^2_{Z_i}$ sua variância. $Cov_{Z_iZ_j}$ representa a covariância entre as variáveis Z_i e Z_j, n é o número de amostras presentes em cada variável. O passo E preenche os dados faltantes por meio de regressões estocásticas a fim de que o passo M possa utilizar o conjunto de equações 11 pra gerar as estimativas dos parâmetros da distribuição. As equações de regressão estocástica calculadas no passo E são obtidas a partir do conjunto 12 de equações:

$$\begin{cases} \beta_1 = \dfrac{Cov_{Z_iZ_j}}{\sigma^2_{Z_j}} \\ \beta_0 = \mu_{Z_i} - \beta_1 \mu_{Z_j} \\ \sigma^2_{Z_i|Z_j} = \sigma^2_{Z_i} - \beta_1^2 \sigma^2_{Z_j} \\ \dot{Z}_i = \beta_0 + \beta_1 Z_j \end{cases} \qquad (12)$$

Onde β_0 representa o coeficiente de interceptação da reta e β_1 o coeficiente angular da reta, $\sigma^2_{Z_i|Z_j}$ é a variância residual da regressão de Z_i a partir de Z_j e $\dot{Z}_i$ é o valor estimado por

meio da equação. O conjunto de equações 12 possibilita o cálculo de valores atualizados de variâncias, covariâncias e médias. É importante ressaltar que ao final, quando atingida a convergência, os valores preenchidos no passo E voltam a não possuir valores, este preenchimento é temporário para que possam ser estimados os parâmetros da distribuição.

Quando no banco de dados multivariado há dados faltantes a mais de uma variável do banco, o passo E do algoritmo deve considerar uma equação de regressão para cada padrão de falta individual. Por exemplo, sendo o banco de dados de três variáveis, Z_1, Z_2 e Z_3, das quais somente a variável Z_1 é completa, o passo E deverá lidar com 3 padrões de falta de dados: somente as variáveis Z_1 e Z_2 possuem amostras, somente as variáveis Z_1 e Z_3 possuem amostras e somente a variável Z_1 possui amostras. Neste caso o passo E do algoritmo necessita das seguintes informações: $\sum Z_1, \sum Z_1^2, \sum Z_2, \sum Z_2^2, \sum Z_3, \sum Z_3^2, \sum Z_1 Z_2, \sum Z_1 Z_3$ e $\sum Z_2 Z_3$. De modo que o passo E do algoritmo irá lidar com os três padrões como blocos de informações, primeiramente irá considerar um padrão obter as informações necessárias para o cálculo da matriz de covariância e seguir para o segundo bloco de informações (presentes no segundo padrão de falta) e assim por diante, até que se obtenham todas as componentes necessárias para a construção do vetor de médias e da matriz de covariâncias no passo M (Enders, 2010).

Estimativa por máxima verossimilhança (Maximum Likelihood Estimation) permite que sejam estimados os parâmetros da distribuição dos dados livres de viés quando o mecanismo de falta é MAR, porém quando o mecanismo de falta for MNAR esta metodologia irá produzir estimativas distorcidas. Além disso, do que foi apresentado até o momento não há nenhuma consideração explicita caso os dados sejam espacialmente correlacionados. Este será o assunto da seção 2.3 deste capítulo que irá apresentar uma metodologia que considera dados espacialmente correlacionados.

2.2 Estimativa bayesiana e inserção múltipla

A estimativa bayesiana consiste em estimar os parâmetros de uma distribuição, chamada de posterior, que descreve a probabilidade relativa para diferentes valores de parâmetros. A abordagem clássica ou freqüentista para estimativa de parâmetros de uma distribuição assume que estes parâmetros são desconhecidos, porém fixos. Ou seja, é assumido o valor que maximiza a função densidade de probabilidade dos dados observados, estimativa por máxima verossimilhança. Por exemplo, considere um experimento com uma moeda onde esta é atirada

para o alto 100 vezes, a probabilidade de obter cara (p_h) ou coroa (p_t) é dada a partir do número de vezes que cada evento ocorre, assim das 100 vezes, se em 75 obtermos cara e 25 coroa as probabilidades correspondentes serão, pela metodologia frequentista, $p_h = \frac{3}{4}$ e $p_t = \frac{1}{4}$.

Já a abordagem bayesiana permite que a probabilidade represente o que se acredita subjetivamente dos dados. No mesmo exemplo da moeda, se subjetivamente crermos que as probabilidades de obter cara ou coroa são iguais, $p_h = \frac{1}{2}$ e $p_t = \frac{1}{2}$, neste caso é assumido que os dados são fixos e que os parâmetros podem assumir possíveis valores.

O paradigma bayesiano lida com os parâmetros da distribuição como variáveis aleatórias que possuem também uma distribuição, tornando a interpretação bem distinta da abordagem clássica.

A análise bayesiana é aplicada em três etapas principais: a primeira, determinar uma distribuição primária, que será referida somente como primária ao longo desta tese, aos parâmetros de interesse. A segunda consiste em sumarizar a evidência dos dados sobre diferentes valores de parâmetros por meio da função densidade de probabilidade, e a terceira, combinar as informações da primária com a função densidade de probabilidade a fim de construir uma distribuição posterior, referida somente de posterior, que descreve a probabilidade relativa de diferentes valores de parâmetros. Obter a posterior é o objetivo principal da análise bayesiana (Enders, 2010).

A primária essencialmente descreve o que se acredita a respeito da probabilidade relativa de diferentes parâmetros antes de qualquer análise exploratória, ou seja, esta etapa exige determinar a média a priori da distribuição, o valor do desvio padrão e o número hipotéticos de pontos amostrais.

A função de probabilidade sumariza as evidências dos dados em relação a diferentes valores de parâmetros, quer dizer, coleta os dados e testa um conjunto de parâmetros possíveis para descrever a distribuição e computa as probabilidades relativas de cada parâmetro.

A posterior é uma composição da primária e função de probabilidade, ponderando cada ponto da função de probabilidade pela primária. Este é o resultado do teorema de Bayes.

Teorema

Sejam A e B_j dois conjuntos. A probabilidade condicional requer que:

$$P(A \cap B) = P(A)P(B|A)$$

Além disso,

$$P(A \cap B) = P(B \cap A) = P(B)P(A|B)$$

Deste modo,

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}$$

onde $P(B|A)$ é a probabilidade condicional de observar o evento B dado que A aconteceu, $P(B)$ é a probabilidade de ocorrer B, $P(A|B)$ é a probabilidade condicional de ocorrer A dado que ocorreu B, $P(A)$ é a probabilidade marginal de A.

Considerando o parâmetro θ da distribuição e o dado Z, o teorema de Bayes é:

$$P(\theta|A) = \frac{P(\theta)P(Z|\theta)}{P(Z)} \qquad (13)$$

A expressão do teorema de Bayes descreve o que é conhecido sobre θ sem conhecimento dos dados por meio de $P(\theta)$, que é a primária, $P(Z|\theta)$ é função probabilidade e $P(Z)$ é a distribuição marginal dos dados, e $P(\theta|A)$ é a posterior que se busca na estimativa bayesiana, essa expressa o que se sabe sobre θ dado o que é conhecido sobre os dados. O teorema de Bayes padroniza o produto da primária e da função de probabilidade pela distribuição marginal dos dados, de modo que a área abaixo da curva da distribuição posterior seja igual a um, ou seja, que a soma das probabilidades descritas pela posterior some um.

A função de probabilidade exerce um papel muito importante no teorema de Bayes, pois é a função por meio da qual os dados modificam o conhecimento a priori (primária de θ) dos parâmetros (Box, 1992).

O teorema de Bayes torna-se bastante atrativo em análise estatística, pois a partir de sua formulação permite que o conhecimento prévio sobre determinado conjunto de amostras seja combinado com novo conhecimento, proveniente da coleta de novas informações, ou seja, o teorema permite a atualização das informações sobre θ à medida que novas amostras são coletadas e novos experimentos realizados (Box, 1992).

A inserção múltipla consiste em três etapas: a etapa de inserção, de análise e finalmente de combinação. A etapa de inserção por sua vez consiste em dois passos: o passo I (inserção) e o passo P (posterior), este último está fortemente alicerçado na análise bayesiana, como será explanado a seguir.

A inserção múltipla é uma alternativa a estimativa por máxima verossimilhança, em que esta não somente realiza uma estimativa dos parâmetros descritivos da distribuição do conjunto

de dados, mas também gera m cenários completos distintos e m desvios padrões, sendo desta forma é possível atestar quanto à incerteza dos valores complementados.

Os m cenários são gerados por meio de uma distribuição bayesiana posterior para os dados observados. O modelo $P(\Phi_{com};\theta)$ implica em uma distribuição condicional para os valores faltantes haja visto os observados, $P(z_{mis}|z_{obs};\theta)$. Considerando que Φ_{com} é o conjunto de dados completo e que este é particionado em amostras presentes, z_{obs}, e amostras faltantes, z_{mis}, os valores complementados poderiam ser diretamente gerados a partir da distribuição condicional para as amostras faltantes caso θ fosse conhecido, porém θ é desconhecido e este deve ser gerado por meio de uma distribuição bayesiana conjunta dos dados observados e de uma variável aleatória R, quando sob mecanismo MNAR. Esta variável R é uma matriz com as mesmas dimensões que a matriz de dados que codifica os valores presentes em 1's e os faltantes em 0's e possui distribuição de falta $P(R|\Phi_{com};\xi)$ e é chamada de "falta" (Schafer, 2003).

Para o caso sob mecanismo de falta MAR θ é obtido a partir de uma distribuição bayesiana que depende somente de z_{obs}. A modelagem de R muitas vezes não é possível, ou se feita, o modelo não é plausível, pois as razões pelas quais os dados são faltantes não constam no banco de dados (Schafer, 2003), portanto a utilização de inserção múltipla se aplica a casos de conjuntos amostrais incompletos sob mecanismo MAR, assim como a estimativa por máxima verossimilhança.

2.2.1 Fase de inserção

Como já mencionado, a inserção múltipla (IM) (Rubin, 1987) é realizada basicamente em três etapas: inserção, análise e combinação. A etapa de inserção gera diversos conjuntos de dados preenchidos com estimativas das amostras faltantes, cada cópia com um valor de estimativa distinto. Esta fase da IM é iterativa e consiste em dois passos: o de inserção (I-step) e o posterior (P-step).

O I-step consiste em construir conjuntos de equações de regressão estocásticas e por meio delas estimar valores faltantes no banco de dados. O P-step, por sua vez, busca realizar estimativas do vetor de médias e da matriz de covariâncias para os cenários completos gerados no I-step. Esta segunda parte está particularmente ligada à análise bayesiana, pois em sua essência busca a posterior do vetor de médias e da matriz de covariâncias.

- I-step:

Sob perspectiva bayesiana, os valores complementados no I-step são sorteios aleatórios de uma distribuição condicional preditiva, dependente dos valores observados, z_{obs}, e das estimativas do vetor de médias e da matriz de covariância, a seguinte equação descreve este procedimento.

$$z_i^* \sim p(z_{mis} | z_{obs}; \boldsymbol{\mu}_{i-1}^*) \quad (14)$$

Onde z_i^* é o valor estimado no instante i, z_{mis} o valor faltante, z_{obs} os observados e $\boldsymbol{\mu}_{i-1}^*$ o vetor de médias e matriz de covariâncias calculados no instante i anterior. De maneira prática, pode-se encarar este sorteio como a adição de um resíduo que possui distribuição normal a cada valor estimado por regressão estocástica, pois a linha de regressão irá interceptar cada distribuição residual em sua média, sendo esta a chamada média condicional, de modo que a adição do valor residual a cada valor predito equivale ao sorteio aleatório da distribuição condicional dependente do valor observado z_{obs} (Enders, 2010).

- P-Step:

O P-step inicia utilizando o conjunto preenchido no I-step precedente para computar um vetor de médias amostral e uma matriz de somas quadráticas e produtos cruzados. A partir destas duas quantidades é possível computar a distribuição posterior da matriz de covariâncias, na forma da equação 15:

$$p(\boldsymbol{\Sigma} | \boldsymbol{\mu}, \boldsymbol{\Phi}) \sim W^{-1}(N-1, \boldsymbol{\Lambda}) \quad (15)$$

onde $p(\boldsymbol{\Sigma} | \boldsymbol{\mu}, \boldsymbol{\Phi})$ é a posterior da matriz de covariância, $\boldsymbol{\Phi}$ é um vetor de variáveis, W^{-1} é a inversa da distribuição de Wishart, N-1 os graus de liberdade e $\boldsymbol{\Lambda}$ a matriz de somas quadráticas e produtos cruzados, $\boldsymbol{\mu}$ é o vetor de média amostral calculada. Tendo a forma da distribuição é possível obter a matriz de covariâncias atualizada por meio de simulação de Monte Carlo, e esta será chamada e matriz de covariância simulada $\boldsymbol{\Sigma}^*$.

Após é calculado o vetor de médias atualizado a partir de sua distribuição posterior, que tem a forma:

$$p(\boldsymbol{\mu} | \boldsymbol{\Phi}, \boldsymbol{\Sigma}) \sim MN(\boldsymbol{\mu}, N^{-1}\boldsymbol{\Sigma}^*) \quad (16)$$

onde $p(\boldsymbol{\mu}|\boldsymbol{\Phi},\Sigma)$ é a posterior do vetor de médias, $\sim MN$ é uma distribuição multivariada normal, $\boldsymbol{\mu}$ o vetor de média amostral do conjunto completo e $\boldsymbol{\Sigma}^*$ a matriz de covariância simulada.

Uma vez que foram calculados os parâmetros atualizados (vetor de médias e matriz de covariância) o algoritmo retorna ao I-step onde calcula um novo conjunto de regressões e gera um novo banco de dados preenchido. Repetir estes dois passos, I e P, m vezes permite gerar m bancos de dados completos. De maneira formal o P-step pode ser descrito pela seguinte equaçao:

$$\Theta_i^* \sim p(\Theta|z_{obs}, z_i^*) \qquad (17)$$

onde Θ_i^* são os parâmetros estimados no instante i, z_{obs} os valores observados e z_i^* os valores complementados no instante i anterior.

Serão demonstradas a seguir as equações 15 e 16 do passo P da inserção múltipla. Por simplificação, será iniciado pela demonstração da posterior para o vetor de médias (equação 16) e em seguida a posterior para matriz de covariâncias (equação 15).

- Distribuição posterior da média:

Para obter a posterior na análise bayesiana, deve-se primeiramente determinar a primária, após a função de probabilidade e a partir da junção destas duas, construir a posterior. A primária é determinada baseada no que se crê sobre a população de amostras, pode-se designar uma distribuição de pesos que favoreça os valores de médias que são mais prováveis, considerando o conhecimento prévio do banco de dados, ou ainda designar uma distribuição onde o analista não assume que um valor de média seja mais provável que outro, portanto distribui pesos iguais a todos os possíveis valores. É importante ter em mente que, jamais o analista deve estar em um estado de completa ignorância a respeito dos dados, porém pode fazer poucas assunções sobre o mesmo (Box, 1973). Primárias desta natureza são chamadas de primárias não informativas.

Para média, no caso univariado onde a variância é conhecida, a primária não informativa é representada por uma linha constante e no multivariado por uma superfície plana.

$$p(\mu|\sigma^2) \propto C \qquad (18)$$

onde C é uma constante. Isto equivale a assumir que a primária é localmente uniforme.

Após a determinação da primária é necessário determinar a função densidade de probabilidade dos dados, que irá sumarizar as informações sobre os parâmetros de distribuição que os dados fornecem. Assumindo uma distribuição normal para a população amostral, a função de probabilidade é expressa da seguinte forma:

$$p(Z|\mu,\sigma^2) = \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-0.5(z_i-\mu)^2}{\sigma^2}} \right\} \qquad (19)$$

onde o termo dentro das chaves é a função densidade de probabilidade para uma distribuição normal. Z é uma variável do conjunto de dados, z_i um valor amostral, μ o valor da média e σ^2 a variância.

Utilizando a primária e a função densidade de probabilidade é possível calcular a posterior da média utilizando o teorema de Bayes. O teorema indica que a altura da posterior é proporcional ao produto da primária com a função de probabilidade, no caso de uma primária não informativa, leva a uma posterior no formato da função de probabilidade, assim:

$$p(\mu|Z,\sigma^2) \sim N\left(\hat{\mu}, \frac{\sigma^2}{N}\right) \qquad (19)$$

onde $\hat{\mu}$ é a média observada nos dados, σ^2 a variância observada nos dados e N o número de amostras no conjunto de dados, ou seja, a distribuição posterior da média é determinada pelos dados, esta é uma consequência de se adotar uma primária não informativa. Para um vetor de médias a primária não informativa será uma superfície plana e a distribuição posterior terá a mesma forma que a função de probabilidade multivariada, resultando na equação 16.

- Distribuição posterior da matriz de covariâncias:

Assim como para o vetor de médias, é necessário determinar uma primária para a matriz de covariâncias a fim de obter sua posterior, porém esta pertence a uma família de distribuições menos conhecidas que a do vetor de médias (multivariada normal), portanto será demonstrado um exemplo univariado, mantendo a média fixa que pode ser extrapolado a situação multivariada.

Será primeiramente considerada a função de probabilidade da variância, a fim de esclarecer a sua primária. Observando uma função de probabilidade normal, para um conjunto com N amostras, a probabilidade amostral é:

$$p(Z|\mu,\sigma^2) \propto \frac{1}{\left(\sigma^2\right)^{N/2}} e^{-0.5\frac{\sum(z_i-\mu)^2}{\sigma^2}} \qquad (20)$$

em que o termo $\sum(z_i-\mu)^2$ é a soma dos desvios quadráticos em torno da média da população que será chamada de SS. A equação 10 pode ser reescrita como:

$$p(Z|\mu,\sigma^2) \propto \frac{1}{(\sigma^2)^{N/2}} e^{-0.5\left(\frac{SS}{\sigma^2}\right)} \qquad (21)$$

Nas equações 19 e 20 está indicada uma proporcionalidade e não uma igualdade, de modo que há um fator de escala a ser definido. A função de probabilidade da equação 20 é o que se chama de distribuição inversa de χ^2, que tem a forma:

$$Inv-\chi^2 \propto \frac{1}{x^{\frac{\nu}{2}+1}} e^{-0.5\left(\frac{S}{x}\right)} \qquad (22)$$

Onde x é uma variável, S um fator de escala que determina o espalhamento da distribuição e ν os graus de liberdade. Como a distribuição χ^2, a sua inversa tem a forma determinada pelos graus de liberdade. Substituindo x por σ^2, S por SS e ν por N, iremos obter a função de probabilidade da equação 21.

Determinada a função de probabilidade, é possível agora com mais clareza entender a forma da primária para a variância. A literatura recomenda que a primária e a função de probabilidade sejam conjugadas e que pertençam à mesma família de distribuições. Assumindo uma $Inv-\chi^2$ de maneira não informativa, ou seja, $N=0$ e $SS=0$, a primária assume a forma:

$$p\left(\sigma^2\right) \propto \frac{1}{\sigma^2} \qquad (23)$$

que decresce a medida que a variância aumenta**.**

Uma vez definidas a primária e a função de probabilidade o último passo da análise bayesiana é determinar a posterior, o teorema de Bayes determina que a posterior é proporcional ao produto da função de probabilidade e primária, assim considerando a média conhecida:

$$posterior \propto primária \times probabilidade$$

$$p\left(\sigma^2|Z,\mu\right) \propto \frac{1}{\sigma^2} \times \frac{1}{\left(\sigma^2\right)^{N/2}} e^{-0.5\left(\frac{SS}{\sigma^2}\right)} = \frac{1}{\sigma^{\frac{N}{2}+1}} e^{-0.5\left(\frac{SS}{\sigma^2}\right)} \qquad (24)$$

Generalizando para o caso onde não somente a variância σ^2 é desconhecida, mas também a média μ, para uma variável aleatória $Z(z_1, z_2, ..., z_N)$ a densidade de probabilidade pode ser escrita como:

$$p\left(Z|\mu,\sigma^2\right)=\left(2\pi\sigma^2\right)^{\frac{-N}{2}} e^{-0.5\frac{\Sigma(z_i-\mu)^2}{\sigma^2}} \qquad (25)$$

De modo que a função de probabilidade pode ser escrita como:

$$L\left(\mu,\sigma^2,Z\right)=\sigma^{\frac{-N}{2}} e^{-0.5\frac{\sum(z_i-\mu)^2}{\sigma^2}}$$

O argumento presente na exponencial pode ser expresso de maneira alternativa:

$$\sum(z_i-\mu)^2=\sum[(z_i-\bar{z})+(\bar{z}-\mu)]^2=\sum(z_i-\bar{z})^2+2(\bar{z}-\mu)\sum(z_i-\bar{z})+N(\bar{z}-\mu)^2$$
$$=\sum(z_i-\bar{z})^2+N(\bar{z}-\mu)^2$$

Substituindo o argumento da exponencial a função de probabilidade se torna:

$$L(\mu,\sigma^2,Z)=\sigma^{\frac{-N}{2}} e^{-0.5\frac{[\sum(z_i-\bar{z})^2+N(\bar{z}-\mu)^2]}{\sigma^2}}$$

Agora, SS que foi definido anteriormente como $\sum(z_i-\mu)^2$ quando μ era conhecido, será redefinido como $\sum(z_i-\bar{z})^2$ e a assim:

$$L(\mu,\sigma^2,Z)=\sigma^{\frac{-N}{2}} e^{-0.5\frac{(SS+N(\bar{z}-\mu)^2}{\sigma^2}} \qquad (26)$$

Definida a função de probabilidade, é necessário definir uma primária que expresse pouco conhecimento sobre os parâmetros. Como não se tem informação sobre os parâmetros da distribuição, a média e a variância podem ser tomadas como independentes (Burkett, 2009). Assim, a densidade de probabilidade conjunta pode ser expressa como o produto das densidades de probabilidades individuais.

$$p(\mu,\sigma^2)=p(\mu)p(\sigma^2) \qquad (27)$$

Assumindo que a densidade de probabilidade da média, $p(\mu)$ é proporcional a um e que a densidade de probabilidade da variância, $p(\sigma^2)$, é proporcional ao inverso da variância, a densidade de probabilidade conjunta será o produto de ambas, de acordo com a equação 27.

$$p(\mu,\sigma^2) \propto \frac{1}{\sigma^2} \qquad (28)$$

Multiplicando a primária pela função de probabilidade é obtida a posterior, de acordo com o teorema de Bayes.

$$p(\mu,\sigma^2|Z) \propto \sigma^{\frac{-N}{2}} e^{-0.5\left[\frac{S+N(\bar{z}-\mu)^2}{\sigma^2}\right]} (\sigma^2)^{-1} \qquad (29)$$

As posteriores marginais podem ser obtidas realizando a integral da posterior conjunta sobre a variância, para posterior marginal da média, e sobre a média para marginal posterior da variância.

$$p(\sigma^2|Z) = \int p(\mu,\sigma^2|Z)d\mu \propto \sigma^{\frac{-(N+1)}{2}} e^{-0.5\left(\frac{SS}{\sigma^2}\right)} \qquad (30)$$

A equação 30 indica que a marginal posterior da variância é aproximadamente uma distribuição χ^2 com N-1 graus de liberdade e com fator de escala SS.

$$p(\sigma^2|Z) \sim Inv-\chi^2(N-1,SS) \qquad (31)$$

Generalizando para a matriz de covariâncias, podem ser substituídos os valores da equação 21 que descreve a função de probabilidade para a variância, como os análogos matriciais, a função de probabilidade para a matriz de covariância será:

$$p(\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{\frac{-N}{2}} e^{-0.5(\mathrm{tr}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}])} \qquad (32)$$

onde $\boldsymbol{\mu}$ é o vetor de médias, $\boldsymbol{\Sigma}$ a matriz de covariância e $\boldsymbol{\Lambda}$ a matriz de somas quadráticas e produtos cruzados. Assim como foi identificado no caso da variância a família de distribuições $Inv-\chi^2$, aqui se identifica uma outra família, a inversa da distribuição de Wishart, W^{-1} (Enders, 2010), que tem a forma:

$$W^{-1} \propto |\boldsymbol{\Sigma}|^{\frac{-(\nu+\kappa+1)}{2}} e^{-0.5(\mathrm{tr}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}])} \qquad (33)$$

onde ν é o grau de liberdade, κ o número de variáveis.

Definida a função de probabilidade, se retorna a primária que como comentado anteriormente, deve ser uma conjugada pertencente à mesma família da função de probabilidade, não informativa. Substituindo $\nu = 0$ e $\mathbf{\Lambda} = 0$, como foi realizado no caso da variância, a primária será:

$$p(\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{\frac{-\kappa+1}{2}} \qquad (34)$$

de modo que $|\mathbf{\Sigma}|$ determina a variação na matriz de covariâncias da população.

Determinadas a primária e função de probabilidade, a posterior é facilmente calculada.

$$p(\mathbf{\Sigma}|\boldsymbol{\mu}, \mathbf{\Phi}) \sim W^{-1}(N-1, \hat{\mathbf{\Lambda}}) \qquad (35)$$

A equação 35 é análoga à equação 31, onde está indicado que a posterior da matriz de covariância é aproximadamente a inversa de Wishart com N-1 graus de liberdade e um fator de escala determinado pela matriz de somas quadráticas e produtos cruzados amostrais, $\hat{\mathbf{\Lambda}}$. Este é o resultado apresentado pela equação 15 no passo P da inserção múltipla. (Enders, 2010).

Na fase de inserção é importante ressaltar que as variáveis selecionadas devem ser o máximo de variáveis disponíveis, pois o uso de inúmeras variáveis no modelo de análise dificilmente irá produzir viés nas estimativas imputadas, já a exclusão de variáveis no modelo de análise potencialmente produzirá viés nas estimativas. Ainda, a exclusão de variáveis pode atenuar as relações entre as variáveis analisadas e não reproduzirá relações complexas quando presentes. Portanto, ao eliminar uma variável da análise o usuário deve garantir que sua exclusão não irá afetar estatisticamente a saída da fase de inserção.

2.2.2 Fase de análise e combinação

A fase de análise da IM tem por objetivo realizar a exploração dos parâmetros e estimativas de cada um dos m cenários gerados na fase de inserção. Isto irá resultar em m estatísticas distintas. A partir destes resultados, a fase de combinação procura colapsar os m cenários em um final. Rubin (1987) define como regra de combinação para as estimativas pontuais de cada realização da inserção múltipla como a média aritmética destas. Portanto,

considerando z_i^* como a i-ésima estimativa pontual, o valor deste ponto estimado, no cenário final será:

$$\bar{z} = \frac{1}{m}\sum_{i=1}^{m} z_i^* \qquad (36)$$

onde $\bar{z}$ é a estimativa pontual combinada final. Note que do ponto de vista frequentista $\bar{z}$ é uma estimativa pontual de um parâmetro fixa da população, enquanto que do ponto de vista bayesiano, esta é a média da distribuição posterior dos dados observados.

A combinação das variâncias de cada cenário, apesar de simplista pela definição de Rubin (1987), não é a média aritmética destas. As variâncias na inserção múltipla possuem duas fontes de flutuação: o erro amostral caso o banco fosse completo e o erro amostral resultante dos dados faltantes. Essas duas fontes são chamadas de variância interna à inserção e variância entre inserções.

A variância interna à inserção, σ_I^2, é dada pela média aritmética das variâncias de cada cenário gerado. Assim:

$$\sigma_I^2 = \frac{1}{m}\sum_{i=1}^{m} \sigma_i^2 \qquad (37)$$

onde σ_i^2 é a variância do i-ésimo cenário gerado, e m o número total de cenários gerados. Esta variância representa a variabilidade amostral caso não houvessem dados faltantes (Enders, 2010).

Já a variância entre inserções, σ_E^2, quantifica a variabilidade dos parâmetros através dos diversos cenários gerados.

$$\sigma_E^2 = \frac{1}{m-1}\sum_{i=1}^{m} (z_i^* - \bar{z})^2 \qquad (38)$$

A variância total, σ_T^2, considerada no cenário final é a composta por estas duas citadas acima, dada por:

$$\sigma_T^2 = \sigma_I^2 + \sigma_E^2 + \frac{\sigma_E^2}{m} \qquad (39)$$

Em que o termo $\frac{\sigma_E^2}{m}$ é um fator de correção dado que o número de inserções realizadas é finito.

2.3 Estimativa bayesiana aplicada à dados geológicos

Até o momento as metodologias não consideram correlação espacial entre variáveis de análise. A correlação espacial não é somente uma característica amostral, mas algo que na mineração se deseja a reprodução. É necessário, quando se lida com um fenômeno geológico a reprodução da sua configuração espacial, para fins de planejamento nas mais diversas áreas da mineração. Motivado por esta deficiência das metodologias estatísticas clássicas, (Doyen et al., 1996; Ren, 2007) propôs uma abordagem bayesiana em que se consideram a correlação espacial entre amostras das variáveis presentes, que após foi utilizada por Banett e Deutsch (2012) evidenciando melhora nos resultados obtidos dos procedimentos geoestatísticos quando utilizando a informação agregada por meio de inserção.

Considere uma variável aleatória Z composta por n amostras, e que é estacionária dentro da região A. Esta será considerada a variável de interesse (primária). Considere ainda que há um conjunto de variáveis secundárias $\mathbf{X}_j$, e que Z e $\mathbf{X}_j$ são multigaussianas.

Como visto anteriormente a estatística bayesiana assume três distribuições: primária, função densidade de probabilidade e posterior, sendo a última o produto das duas primeiras. Em sua tese Ren (2007) utiliza os resultados de krigagem simples da variável de interesse Z como parâmetros da distribuição primária, ou seja:

$$\bar{z}_p(u) = \sum_{i=1}^{n} \lambda_i \cdot z_p(u_i) \qquad (40)$$

Onde $\bar{z}_p$ é a média da distribuição primária, u o local de estimativa, u_i o local onde a variável de interessa está amostrada e λ_i os pesos de krigagem simples calculados a partir da equação 41.

$$\sum_{i=1}^{n} C(u_i - u_k) = C(u - u_k); k = 1,...,k \qquad (41)$$

$C(u_i - u_k)$ é a covariância entre a variável primária no local u_i e a variável primária no local u_k, e $C(u - u_k)$ a covariância entre a primária no local a ser estimado e a primária no local u_k.

A variância da distribuição é obtida a partir da equação 42.

$$\sigma_p^2(u) = 1 - \sum_{i=1}^{n} \lambda_i C(u, u_i) \qquad (42)$$

Com estes foi parâmetros a distribuição primária, que é uma gaussiana está completamente definida.

Se os dados secundários $\mathbf{X}_j$ estiverem disponíveis em toda a área de estacionariedade da variável primária, os resultados de estimativa utilizando os dados secundários dão origem à função densidade de probabilidade, segunda parte da análise bayesiana.

$$\bar{z}_L(u) = \sum_{i=1}^{m} \lambda_i \cdot x_i(u) \qquad (43)$$

em que $x_i(u)$ é o valor da amostra secundária no local de estimativa u e $\bar{z}_L$ a média da função densidade de probabilidade. Os pesos são calculados pela equação 44.

$$\sum_{j=1}^{m} \lambda_j \cdot \rho_{j,k} = \rho_{i,0} \qquad (44)$$

$\rho_{j,k}$ é a correlação entre duas variáveis secundárias distintas, e $\rho_{i,0}$ é o valor entre a variável primária e secundária. A variância da função densidade de probabilidade é dada por:

$$\sigma_L^2(u) = 1 - \sum_{j=1}^{m} \lambda_j \cdot \rho_{j,0} \qquad (45)$$

Definidas as duas distribuições que compõe a posterior, pode-se calcular esta última por meio do produto das duas primeiras:

$$P\{z(u)|X(u), z(u_1),..., z(u_n)\} \propto f\{X(u)|z(u)\} P\{z(u)|z(u_1),..., z(u_n)\} \qquad (46)$$

O termo $P\{z(u)|z(u_1),..., z(u_n)\}$ na equação 46 é a distribuição primária calculada, condicionada somente pela variável de interesse, lembrando que está é uma distribuição gaussiana com média e variância de krigagem simples, portanto:

$$P\{z(u)|z(u_1),..., z(u_n)\} \propto \exp\left[\frac{\left(z(u) - \bar{z}_p(u)\right)^2}{2\sigma_p^2(u)}\right] \qquad (47)$$

O símbolo $\propto$ indica que as constantes de proporcionalidade foram omitidas. Expandindo os termos quadrático, $\left(z(u)-\bar{z}_p(u)\right)^2$:

$$\exp\left[\frac{\left(z(u)-\bar{z}_p(u)\right)^2}{2\sigma_p^2(u)}\right]=\exp\left[\frac{z(u)^2}{2\sigma_p^2(u)}+\frac{z(u)\bar{z}_p(u)}{\sigma_p^2(u)}-\frac{\bar{z}_p^2(u)}{2\sigma_p^2(u)}\right]$$

Assim, eliminando as constantes independentes de $z(u)$:

$$P\left\{z(u)\middle|z(u_1),...,z(u_n)\right\}\propto\exp\left[\frac{z(u)^2}{2\sigma_p^2(u)}+\frac{z(u)\bar{z}_p(u)}{\sigma_p^2(u)}\right] \qquad (48)$$

O termo $f\left\{X(u)\middle|z(u)\right\}$ da equação 46 representa a função densidade de probabilidade, calculada utilizando as variáveis secundárias colocadas. No contexto multigaussiano essa função densidade de probabilidade é parametrizada pelo vetor de média condicional, $E\left\{X(u)\middle|z(u)\right\}=\boldsymbol{\rho}z(u)$, e pela matriz de covariância condicional $\boldsymbol{\Sigma}$. $\boldsymbol{\rho}$ é o vetor de correlação entre a variável de interesse e as secundárias do banco de dados. Assim, a matriz de covariância pode ser escrita em termos dos coeficientes de correlação.

$$\begin{bmatrix}\rho_{11} & \cdots & \rho_{1n}\\ \vdots & \ddots & \vdots\\ \rho_{n1} & \cdots & \rho_{nn}\end{bmatrix}-\begin{bmatrix}\rho_{10}\rho_{10} & \cdots & \rho_{10}\rho_{n0}\\ \vdots & \ddots & \vdots\\ \rho_{n0}\rho_{10} & \cdots & \rho_{n0}\rho_{n0}\end{bmatrix}=\boldsymbol{\rho}_{\mathbf{ij}}-\boldsymbol{\rho}\boldsymbol{\rho}^{\mathbf{T}}$$

Como os dados secundários são conhecidos, $f\left\{X(u)\middle|z(u)\right\}$ será uma função de $z(u)$ somente (Ren, 2007).

$$f\left\{X(u)\middle|z(u)\right\}\propto\exp\left[-\frac{1}{2}\left(X(u)-\boldsymbol{\rho}z(u)\right)^T\boldsymbol{\Sigma}^{-1}\left(X(u)-\boldsymbol{\rho}z(u)\right)\right] \qquad (49)$$

Onde

$$\begin{aligned}&\exp\left[-\frac{1}{2}\left(X(u)-\boldsymbol{\rho}z(u)\right)^T\boldsymbol{\Sigma}^{-1}\left(X(u)-\boldsymbol{\rho}z(u)\right)\right]\\&=\exp\left[-\frac{1}{2}\left(X^T(u)-\boldsymbol{\rho}^Tz(u)\right)\boldsymbol{\Sigma}^{-1}\left(X(u)-\boldsymbol{\rho}z(u)\right)\right]\\&=\left[-\frac{1}{2}\left(X^T(u)\boldsymbol{\Sigma}^{-1}X(u)-2\boldsymbol{\rho}^T\boldsymbol{\Sigma}^{-1}X(u)z(u)+\boldsymbol{\rho}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\rho}z(u)^2\right)\right]\end{aligned}$$

Omitindo as constantes independentes de $z(u)$. Assim:

$$f\left\{X(u)\middle|z(u)\right\}\propto\left(\boldsymbol{\rho}^{T}\boldsymbol{\Sigma}^{-1}X(u)z(u)-\frac{1}{2}\boldsymbol{\rho}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\rho}z(u)^{2}\right) \qquad (50)$$

Ao multiplicar as equações 50 e 48 é gerada a distribuição posterior.

$$P\left\{z(u)\middle|X(u),z(u_1),...,z(u_n)\right\}\propto\exp\left[-\frac{1}{2}\left(\boldsymbol{\rho}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\rho}+\frac{1}{\sigma_p^2(u)}\right)z(u)^2+\left(X^{T}(u)\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\rho}+\frac{\bar{z}_p(u)}{\sigma_p^2(u)}\right)z(u)\right] \quad (51)$$

O produto da distribuição primária pela função densidade de probabilidade é proporcional a uma exponencial da soma de duas constantes, $\exp[-Ax^2+Bx]$, a partir das constantes A e B é possível obter a média e variância de uma kernel gaussiana:

$$\bar{z}_U(u)=\frac{B}{2A}=\frac{\boldsymbol{\rho}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}X(u)+\dfrac{\bar{z}_p(u)}{\sigma_p^2(u)}}{\boldsymbol{\rho}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\rho}+\dfrac{1}{\sigma_p^2(u)}}=\frac{\dfrac{\boldsymbol{\rho}^{\mathbf{T}}\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}X(u)}{\left(\boldsymbol{\rho}_{\mathbf{ij}}-\boldsymbol{\rho}\boldsymbol{\rho}^{\mathbf{T}}\right)\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}}+\dfrac{\bar{z}_p(u)}{\sigma_p^2(u)}}{\dfrac{\boldsymbol{\rho}^{\mathbf{T}}\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}\boldsymbol{\rho}}{\left(\boldsymbol{\rho}_{\mathbf{ij}}-\boldsymbol{\rho}\boldsymbol{\rho}^{\mathbf{T}}\right)\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}}+\dfrac{1}{\sigma_p^2(u)}} \qquad (51)$$

$$=\frac{\boldsymbol{\rho}^{\mathbf{T}}\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}X(u)\sigma_p^2(u)+\bar{z}_p(u)(1-\boldsymbol{\rho}\boldsymbol{\rho}_{\mathbf{ij}}\boldsymbol{\rho}^{\mathbf{T}})}{\boldsymbol{\rho}^{\mathbf{T}}\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}\sigma_p^2(u)+(1-\boldsymbol{\rho}\boldsymbol{\rho}_{\mathbf{ij}}\boldsymbol{\rho}^{\mathbf{T}})}$$

Sabe-se da equação 35 que os pesos utilizados no cálculo da média e variância da função densidade de probabilidade podem ser expressos em termos da correlação, de modo que:

$$\bar{z}_L(u)=\boldsymbol{\rho}^{\mathbf{T}}\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}X(u) \qquad \text{e} \qquad \sigma_L^2(u)=1-\boldsymbol{\rho}^{\mathbf{T}}\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}\boldsymbol{\rho}$$

Logo, a expressão para a média da distribuição posterior pode ser escrita como:

$$\bar{z}_U(u)=\frac{\boldsymbol{\rho}^{\mathbf{T}}\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}X(u)\sigma_p^2(u)+\bar{z}_p(u)(1-\boldsymbol{\rho}\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}\boldsymbol{\rho}^{\mathbf{T}})}{\boldsymbol{\rho}\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}\boldsymbol{\rho}^{\mathbf{T}}\sigma_p^2(u)+(1-\boldsymbol{\rho}\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}\boldsymbol{\rho}^{\mathbf{T}})}=\frac{\bar{z}_L(u)\sigma_p^2(u)+\bar{z}_p(u)\sigma_L^2(u)}{\sigma_p^2(u)-\sigma_L^2(u)\sigma_p^2(u)+\sigma_L^2(u)} \quad (52)$$

E a expressão para variância da distribuição posterior é:

$$\sigma_U^2(u)=\frac{1}{2A}=\frac{1}{\boldsymbol{\rho}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\rho}+\dfrac{1}{\sigma_p^2(u)}}=\frac{(1-\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}\boldsymbol{\rho}^{\mathbf{T}})\sigma_p^2(u)}{\boldsymbol{\rho}^{\mathbf{T}}\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}\sigma_p^2(u)+(1-\boldsymbol{\rho}\boldsymbol{\rho}_{\mathbf{ij}}^{-\mathbf{1}}\boldsymbol{\rho}^{\mathbf{T}})}=\frac{\sigma_L^2(u)\sigma_p^2(u)}{\sigma_p^2(u)-\sigma_p^2(u)\sigma_L^2(u)+\sigma_L^2(u)} \quad (53)$$

A figura 1 ilustra o funcionamento da atualização bayesiana.

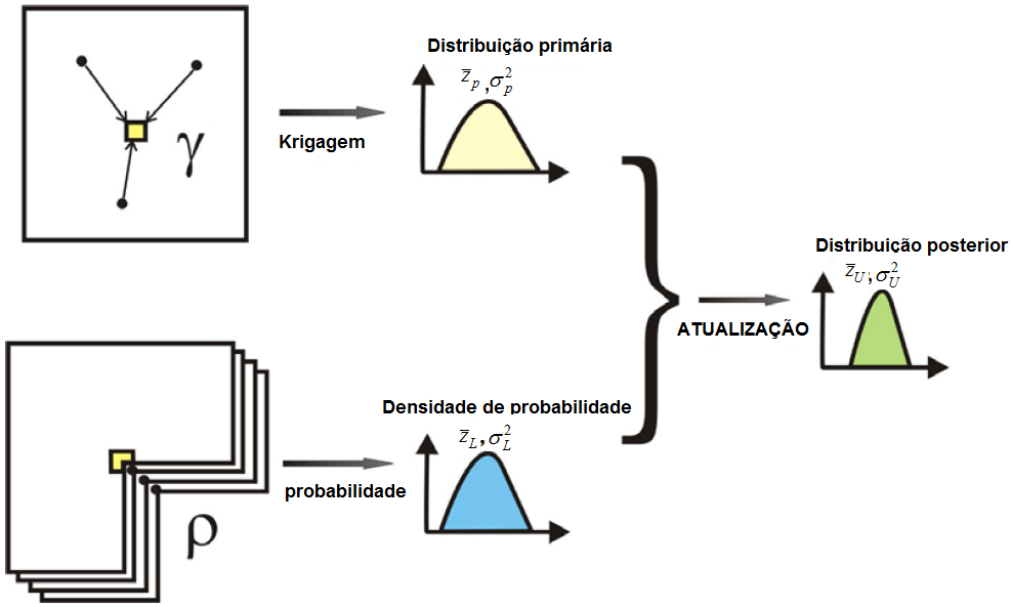

Figura 1 - Diagrama da atualização bayesiana. O quadrado amarelo representa o local a ser complementado e os pontos pretos as amostras disponíveis. Figura adaptada (Ren, 2007)

Desta maneira a distribuição posterior fica completamente definida, o que permite a inserção de dados via simulação estocástica, amostrando a distribuição posterior, onde são mantidas constantes as amostras existentes no banco de dados e os valores complementados terão um grau de incerteza baseado na distribuição posterior (Barnett e Deutsch, 2012). A metodologia proposta por Ren (2007) e depois utilizada por Banett e Deutsch (2012) assume a multi-gaussianidade dos dados, seguramente os dados não terão suas distribuições condicionais e marginais gaussianas, portanto é necessário aplicar, previamente a técnica de atualização bayesiana a transformação *nscore* (Verly, 1983; Deutsch e Journel, 1998), que levará as distribuições originais dos dados às distribuições gaussianas. Após a aplicação da técnica proposta cada distribuição deve ser retrotransformada às unidades originais dos dados.

2.4 Metodologias MNAR

Quando os dados são faltantes não aleatórios, MNAR, é sabido que os valores não amostrados carregam informações importantes sobre os parâmetros do modelo da população, de modo que ignorar este mecanismo de falta acarreta em estimativas distorcidas destes parâmetros. Os modelos MNAR mitigam o viés incorporando uma descrição da propensão de falta de dados. Dentre as metodologias para tratar de dados faltantes não aleatórios estão: modelo de seleção e modelo de mistura de padrões. Estas duas metodologias incorporam a propensão de falta de dados de maneiras bastante distintas em prática. Por exemplo, o modelo de seleção incorpora uma equação de regressão que prevê a probabilidade de falta de dados

enquanto o modelo de mistura de padrões estratifica o conjunto de dados em padrões de falta e estima os parâmetros individualmente para cada padrão. Cada modelo será explorado nos itens subsequentes.

2.4.1 Modelos de seleção

Considerando que cada variável do conjunto de dados possui um par de amostras: Z e R. Z é um valor que pode ou não ser observado, por exemplo, o teor do minério amostrado e R é um código binário que representa o fato da variável ter sido amostrada ou não. Assim, R=1 se Z foi amostrado e R=0 se Z é amostra faltante (Rubin, 1976). Quando se trabalha sob mecanismo MNAR, se sabe que há um modelo desconhecido que governa o fato de R assumir o valor um ou zero.

Uma vez que os dados amostrados dependem dos faltantes e vice-versa, se trabalha com uma distribuição conjunta que descreve a ocorrência mutua de diferentes valores amostrais e de dados faltantes, $p(Z,R)$.

O modelo de seleção (Heckmann, 1979) propõe uma fatoração desta distribuição conjunta em duas distribuições componentes:

$$p(Z,R) = p(R|Z)\,p(Z) \qquad (54)$$

onde $p(R|Z)$ uma distribuição condicional de falta que descreve a probabilidade de uma variável com determinado valor ser faltante ou não, e $p(Z)$ é a distribuição marginal dos dados que descreve a probabilidade de se obter valores distintos de Z.

Em seu modelo, Heckmann (1979) propõe tratar esta fatoração em duas partes, combinando um modelo de regressão, que aqui será chamado de fundamental, pois é o modelo de regressão que seria estimado caso não houvesse dados faltantes no conjunto. O segundo seria um modelo de regressão que prevê a probabilidade de resposta de uma variável. Esta última parte define a probabilidade de ocorrerem faltantes como uma variável latente, F, com distribuição normal. Casos que estão acima de um determinado valor de corte estão presentes, casos abaixo deste valor de corte são faltantes.

Para ilustrar o que foi descrito acima, suponha um conjunto de dados com as variáveis Ni (ppm), Co (ppm) e uma terceira variável Zn (ppm) que está completamente amostrada. Ni (ppm) é a variável faltante como nos exemplos bivariados já apresentados nesta tese.

O modelo de regressão fundamental seria:

$$Ni = \beta_1 + \beta_0 Co + \varepsilon \qquad (55)$$

onde β_0 e β_1 são os coeficientes de regressão e ε o termo residual de regressão. Esta equação corresponde a distribuição marginal de Z. O modelo de regressão que prevê a probabilidade de falta seria:

$$F = \alpha_1 + \alpha_0 Zn + \xi \qquad (56)$$

onde α_0 e α_1 são os coeficientes de regressão e ξ o termo residual. Esta equação corresponde à distribuição condicional de falta. Como F é uma variável latente e, portanto, completamente faltante, é usada então a variável binária R que descreve se o valor é presente ou não no conjunto de dados, que servirá como manifesto da variável latente F. Para construir um modelo de regressão em que a variável de saída é binária pode-se utilizar um modelo de regressão logístico ou probit, apesar dos dois serem bastante semelhantes, a literatura recomenda o modelo probit.

O modelo probit (probability unit) teve seu conceito desenvolvido por Bliss (1934) em um artigo onde Bliss trata de dados de porcentagem de uma peste morta por um determinado pesticida, e ele propõe transformar esta porcentagem em uma unidade de probabilidade. O modelo é uma regressão em que variável dependente pode assumir dois valores somente (binária) e o objetivo é estimar a probabilidade que uma observação tem de pertencer a uma determinada classe ou de estar acima de um valor de corte.

O modelo probit é recomendado para o caso aqui tratado, pois este está associado à distribuição normal padrão. O modelo utiliza a distribuição cumulativa normal padrão (figura 2) para gerar as probabilidades estimadas.

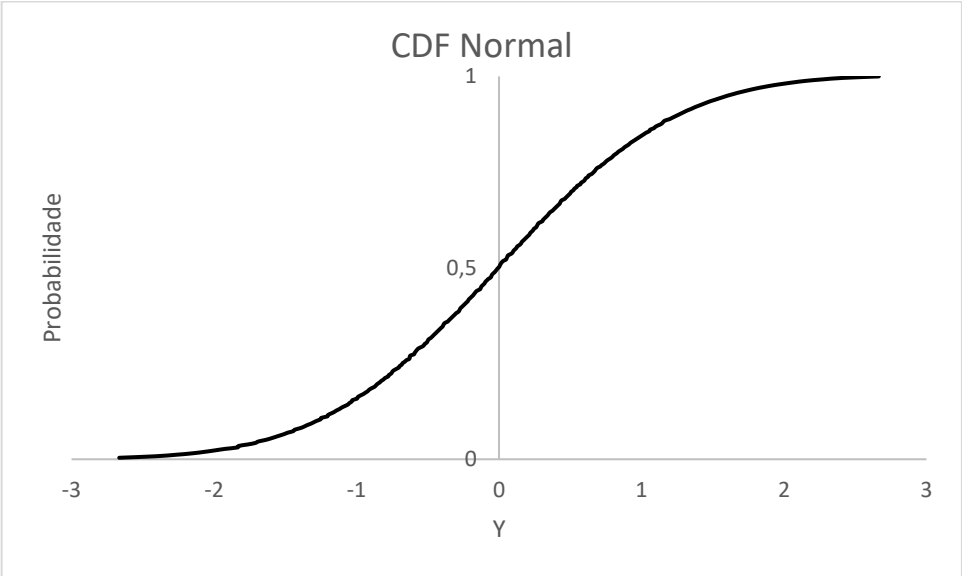

Figura 2- Distribuição cumulativa normal padrão (CDF)

Voltando ao exemplo anterior onde Ni (ppm), Co (ppm) e Zn (ppm) formam um conjunto de dados em que Ni (ppm) é a variável com amostras faltantes, a regressão para estimar a probabilidade de faltantes (equação 56) reescrita em termos da variável binária R é:

$$p(R=1|Zn) = \Phi[\alpha_1 + \alpha_0(Zn)] \qquad (57)$$

onde Φ representa a distribuição cumulativa normal padrão e α_0 e α_1 os coeficientes de regressão.

Para estimar os valores de probabilidade é necessário estimar os parâmetros da regressão, o que pode ser realizado por meio de estimativa por máxima verossimilhança, descrito no capítulo 2 desta tese. Uma vez estimados os parâmetros, pode-se calcular um valor de g para o termo entre colchetes da equação 57, e após transformar este valor g por meio de Φ, o que resultará no valor de probabilidade de falta que se busca estimar.

A partir deste modelo, se tem a distribuição marginal dos dados (regressão fundamental) e a distribuição condicional de falta (regressão probit) de modo que é possível obter a distribuição conjunta de dados e de falta de dados. Para realizar a inserção de dados podem ser feitos sorteios de valores desta distribuição conjunta através de simulação de monte carlo e amostrador de Gibbs entre outros para obter os valores a serem complementados.

O modelo de seleção, porém apresenta algumas limitações que devem ser destacadas. A equação de regressão, que prevê a probabilidade de falta de dados, é a maneira que o modelo de seleção elimina o viés das estimativas quando o mecanismo MNAR é presente no conjunto de dados. Contudo, quando as variáveis explanatórias do modelo probit são correlacionadas às variáveis explanatórias do modelo fundamental, as probabilidades estimadas serão também correlacionadas às variáveis do modelo fundamental, o que introduz erro amostral nas estimativas geradas (Puhani, 2000; Enders, 2010). Além disso, o modelo de seleção fundamenta-se em hipóteses de distribuições normais, de modo que caso os dados não atendam a tais hipóteses isto também fará com que o modelo produza estimativas distorcidas. Há estudos que sugerem que a performance do modelo de seleção é satisfatória, quando as hipóteses que este assume são cumpridas; porém, estas são fortemente instáveis (Hartmann, 1991; Nwata, 1993) de maneira que não há como avaliar de forma prática a qualidade do modelo.

2.4.2 Modelo de mistura de padrões

O modelo de mistura de padrões integra uma distribuição que descreve a falta de dados na análise estatística, portanto na estimativa dos parâmetros de distribuição da população. O

método realiza esta tarefa formando subgrupos de casos de dados faltantes que compartilham o mesmo padrão de falta e estima a distribuição da população de cada subgrupo individual. Os subgrupos serão chamados aqui de padrões específicos. Os modelos estimados para cada padrão específico individualmente não são de interesse, mas sim a média dos parâmetros estimados, que levará a um modelo único resultante que considera o mecanismo MNAR e por consequência elimina o viés adicionado às estimativas quando o mecanismo MAR é violado.

Considerando a distribuição conjunta de dados e de falta de dados, o modelo de mistura de padrões fatora de maneira distinta ao modelo de seleção:

$$p(Z,R) = p(Z|R)\,p(R) \qquad (58)$$

de maneira que a distribuição de Z está condicionada ao mecanismo de falta R. Neste contexto, cada modelo de padrão específico corresponde a distribuição condicional $p(Z|R)$ e as proporções de cada padrão de falta correspondem a distribuição marginal de $p(R)$. Os padrões de falta podem ser monotônicos, onde as variáveis estão dispostas em uma ordem tal que, se a variável é faltante para Z_j é faltante para Z_{j+1}. Ou ainda, o padrão pode ser não monotônico, onde a falta não é ordenada entre variáveis, a figura 3 ilustra os dois casos.

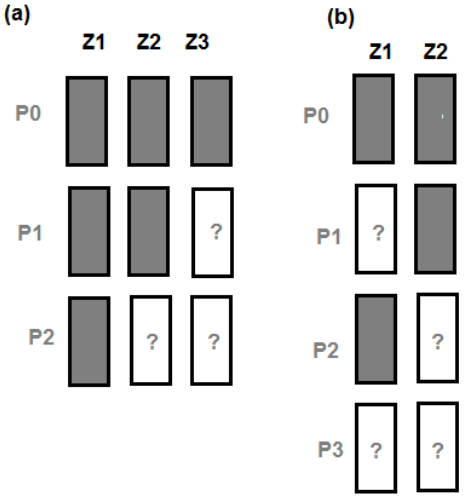

Figura 3 - Padrões de falta. (a) padrão monotônico; (b) padrão não monotônico. Figura adaptada (Litlle, 1993)

Na figura 3, estão ilustrados os padrões monotônicos e não monotônicos, onde P_j representa um padrão especifico de falta, por exemplo, na figura 5b, se tem que P_0 é o caso completo, P_3 completamente faltante e P_1 e P_2 casos onde somente Z_2 e Z_1 estão amostrados respectivamente. A distribuição conjunta para cada padrão específico é o produto da distribuição marginal e da distribuição condicional das variáveis faltantes dadas as observadas. A distribuição condicional para cada padrão específico será chamada de distribuição de variável

faltante (VF) e para identificar os parâmetros de cada VF são utilizadas restrições, chamadas de restrições de variáveis faltantes. Por exemplo, considerando o banco de dados de Ni (ppm) e Co (ppm) apresentado na tabela 1, são identificados dois padrões de falta: casos onde há tanto a variável Ni (ppm) quanto Co (ppm) e casos onde há somente Co (ppm).

Tabela 1 - conjunto de Co (ppm), Ni (ppm) e Ni (ppm) incompleto

Co (ppm)	Ni (ppm)	Ni (ppm) incompleto
9,32	21.32	----
10	29.72	29,72
10,6	21.4	---
11,92	29.72	----
16,32	26.2	----
3,58	22.04	22,04
15,08	21.76	21,76
4,2	9.72	----
4,52	11.32	----
12,08	16.88	----
9,84	21.64	21,64
7,56	15.72	15,72
2,368	7.08	7,08
8,84	17.4	17,4
11,96	23.76	23,76

Considerando que as duas variáveis possuam uma relação linear e que o objetivo seja estimar a média de Ni (ppm), que é a variável incompleta, a média desta será obtida por meio de uma regressão linear do tipo:

$$\mu_{Ni(ppm)} = \beta_0 + \beta_1 \mu_{Co(ppm)} \qquad (59)$$

onde β_0 e β_1 são os coeficientes de regressão de Ni (ppm) em função de Co (ppm). Porém, os parâmetros β_0 e β_1 não são estimáveis, pois, para fazê-lo corretamente deveria se ter a amostragem de Ni (ppm) completa. Desse modo, os coeficientes de regressão são substituídos pelos coeficientes que são estimáveis, do subconjunto onde há tanto Ni (ppm) quanto Co (ppm) (amostragem isotópica), α_0 e α_1. Considerando os casos onde Ni (ppm) e Co (ppm) estão presentes, temos uma equação de regressão do tipo:

$$Ni(ppm) = 0{,}9267\,Co(ppm) + 11{,}879 \qquad (60)$$

onde $\alpha_0 = 11{,}879$ e $\alpha_1 = 0{,}9276$, são os coeficientes de regressão para o subconjunto isotópico. Utilizando estes coeficientes, é possível estimar a média de Ni (ppm) onde este é faltante.

$$\mu^{(f)}_{Ni(ppm)} = 11{,}879 + 0{,}9267\,(\mu^{(f)}_{Co(ppm)}) \qquad (61)$$

Na equação 61, $\mu^{(f)}_{Ni(ppm)}$ representa os casos onde Ni (ppm) é faltante e $\mu^{(f)}_{Co(ppm)}$ a média de Co (ppm) considerando somente as amostras em que Ni (ppm) é faltante. O valor para$\mu^{(f)}_{Co\,(ppm)} =$ 8,64 (ppm) leva a um valor de média de Ni (ppm) nos casos faltantes $\mu^{(f)}_{Ni(ppm)} = 19{,}88$ (ppm). A média de Ni (ppm) considerando o subconjunto isotópico é $\mu^{(o)}_{Ni(ppm)}$= 19,89 (ppm). A média estimada final de Ni (ppm), $\mu^{E}_{Ni(ppm)}$, é a média ponderada entre o valor obtido pela regressão em que Ni (ppm) é faltante, $\mu^{(f)}_{Ni(ppm)}$, e o valor calculado para o subconjunto isotópico, $\mu^{(o)}_{Ni(ppm)}$. Cada valor é considerado de acordo com a proporção de falta, por exemplo, 47% das amostras são faltantes de Ni (ppm) e 53% do total de amostras possuem valores para Ni (ppm). Portanto, se tem o valor $\mu^{E}_{Ni(ppm)} = 0{,}47(19{,}88) + 0{,}53(19{,}89) = 19{,}88$ o que leva a um desvio relativo de -0,81% do valor da média do conjunto completo em que a média é 19,71 ppm. No mesmo exemplo, se aplicada somente a regressão linear (equação 61) aos dados sem a separação em padrões de falta, a média obtida para a variável Ni (ppm) é de 20,40 (ppm), com desvio relativo de 3,55% quando comparada à média do conjunto completo real.

A abordagem adotada no exemplo ilustrativo do modelo de mistura de padrões chama-se CASO-COMPLETO VARIÁVEL FALTANTE (complete case missing variable - CCMV), onde os parâmetros de distribuições de cada padrão são igualados aos parâmetros da distribuição de caso completo e isto pode ser realizado para quantos forem os padrões identificados (Little, 1993). Outras restrições de variáveis faltantes podem ser adotadas, por exemplo, ao invés de igualar os parâmetros dos padrões de falta ao caso completo, podem-se igualar parâmetros em subgrupos similares, chamada de **restrição de vizinho próximo.**

O modelo de mistura de padrões, porém também apresenta limitações quanto a sua eficiência, pois é preciso que o usuário atribua valores a parâmetros que em essência são inestimáveis, devido a incompletude do conjunto de dados. Caso os valores atribuídos estejam corretos o modelo é capaz de eliminar o viés presente devido ao mecanismo de falta atuante, todavia, caso os valores atribuídos estejam incorretos este pode gerar valores ainda mais enviesados que os obtidos por meio de modelo MAR (Enders, 2010). Os valores assumidos e

hipóteses adotadas no modelo de mistura de padrões não são passiveis de teste, portanto a atribuição de valores a tais parâmetros deve ser evitada.

2.4.3 Transformações fixas

Outra abordagem que possui apelo devido à sua praticidade de aplicação, é a que consiste em transformações fixas. Segundo Rubin (1987), quando o mecanismo de falta é não aleatório, existe uma diferença sistemática entre os dados observados e os dados faltantes. Como não há maneira de estimar diretamente o viés gerado pela discrepância amostral, o modelo será sensível às hipóteses adotadas em relação as similaridades entre observados e não observados.

Portanto, Rubin (1987) sugere que modelos sob diversas hipóteses sejam gerados e que o analista, em posse do conhecimento empírico e prático, faça juízo de valores a respeito dos méritos dos modelos. Deste modo, os modelos adotados devem ser facilmente comunicáveis entre si, mesmo que modelos mais sofisticados possivelmente sejam representações mais acuradas da realidade, modificações descritas à modelos adotados assumindo mecanismo MAR tendem a ser aplicados devido à sua facilidade.

Os modelos MNAR mitigam o viés incorporando uma descrição da propensão de falta de dados. Quando o banco de dados viola a hipótese de MAR, ou seja, quando a probabilidade de a amostra ser faltante depende do seu próprio valor, as metodologias descritas anteriormente podem apresentar viés aos resultados obtidos. Para contornar esta limitação do método de estimativa por máxima verossimilhança, ou das metodologias baseadas em inserção múltipla, é preciso que seja incorporada ao modelo dos dados uma componente que descreva a probabilidade de falta amostral. Integrar tal componente torna-se um desafio, pois não há acesso aos valores faltantes.

Rubin (1987) propõe que transformações fixas sejam aplicadas às estimativas obtidas por meio de modelos MAR, tornando-as assim em saídas MNAR. Uma vez que os cenários complementados foram gerados (via estimativa por máxima verossimilhança ou inserção múltipla ou ainda, por atualização bayesiana assumindo mecanismo MAR), seja adicionada uma constante aos resultados que compense uma super ou subestimativa. Por exemplo, se é sabido que determinado minério não foi amostrado devido ao seu baixo teor, então todas as suas amostras terão valores altos o que levará a resultados superestimados. Assim ao final da inserção, o analista pode adicionar uma constante aos resultados que baixe o valor das estimativas e observar a sensitividade estatística do conjunto construído. Uma das transformações fixas proposta por Rubin (1987) é que cada valor complementado assumindo

um mecanismo MAR, seja transformado em um valor MNAR por adição ou subtração de 20% do valor complementado. A transformação fixa é atrativa devido a sua simplicidade de aplicação e fácil comunicação entre o modelo MAR e MNAR. A escolha da constante; porém, é arbitrária.

A análise de sensitividade busca avaliar a variabilidade do conjunto de dados em diversos cenários com inúmeras condições. Desta maneira, a constante utilizada pode ser modificada. Cohen (1988) alvitra que a constante utilizada seja determinada pelo desvio padrão dos cenários complementados.

Portanto, considerando as metodologias apresentadas nesta seção, a abordagem proposta por Rubin (1987) foi adotada no presente trabalho. É proposto que a transformação fixa seja determinada pelo erro máximo obtido em cada cenário complementado assumindo um modelo MAR para os dados. Como a abordagem apresentada permite fácil comunicação entre modelos gerados a partir de diferentes hipóteses, o produto gerado utilizando a transformação fixa sugerida nesta tese é confrontado com os modelos obtidos por meio de atualização bayesiana e por meio da transformação fixa de Rubin (1987), de 20% do valor complementado. A metodologia é descrita em detalhes no capítulo subsequente.

CAPITULO 3

Metodologia

Neste capitulo, é apresentada metodologia aplicada para inserção de dados faltantes não aleatórios (MNAR).

3.1 Transformações fixas às inserções obtidas por meio de atualização bayesiana (AB)

No contexto geoestatístico, é necessária a integração de toda informação disponível, considerando a estruturação espacial dos atributos. Portanto, Doyen (1996) propôs uma simplificação da cokrigagem colocada onde o modelamento de todas as covariâncias diretas e cruzadas é evitada e a estimativa cokrigada é obtida diretamente por meio da regra Bayes. Em 2007, Ren aplicou as ideias de Doyen ao problema de inserção de amostras faltantes no conjunto de dados, como apresentado na seção 2.3 do capítulo 2.

A metodologia supracitada, mostrou-se adequada quando aplicada a um conjunto de dados incompleto sob mecanismo de falta aleatório, porém quando o mecanismo de falta é não aleatório tal metodologia, assim como estimativa por máxima verossimilhança e inserção múltipla, pode apresentar viés aos resultados obtidos. Como descrito na seção 3.5 do capitulo 3, uma abordagem direta é proposta por Rubin (1987), onde são aplicadas transformações fixas às saídas dos modelos gerados assumindo mecanismo MAR, a fim de obter valores complementados sob o mecanismo MNAR.

Neste estudo, é proposto que o processo de inserção seja realizado por meio de atualização bayesiana, de modo que o modelo é gerado admitindo o mecanismo MAR e após uma transformação fixa é aplicada aos valores complementados para que estes sejam resultantes de um mecanismo MNAR.

A transformação aqui proposta se dá pela adição de uma constante igual ao erro relativo máximo obtido pelo modelo MAR, de acordo com as equações 62 e 63:

$$Err_j = \frac{z(u_i) - z_j^*(u_i)}{z(u_i)} \qquad (62)$$

$$z_{j'}^*(u_i) = z_j^*(u_i) + z_j^*(u_i) * Err_j \qquad (63)$$

onde Err_j é o erro relativo máximo obtido no j-ésimo cenário, $z(u_i)$ é o valor observado no local u_i, $z_j^*(u_i)$ é o valor complementado no local u_i admitindo mecanismo MAR e $z_{j'}^*(u_i)$ é o valor MNAR resultante da transformação fixa.

É importante ressaltar que o cálculo do erro relativo é obtido por meio de um conjunto, aqui referido como conjunto de calibração, que consiste no banco de dados original de dados menos 20% das amostras da variável de interesse (ou faltante). Tais amostras são removidas aleatoriamente a fim de que não seja gerado viés adicional ao grupo de dados em análise. Desta maneira, o conjunto de calibração é submetido à inserção por meio de atualização bayesiana assumindo o mecanismo MAR, e os valores que foram removidos podem ser comparados aos valores complementados, de modo que o erro relativo do processo de inserção pode ser diretamente calculado. Nesta etapa são gerados tantos cenários complementados para o conjunto de calibração quanto se desejam obter para o banco de dados original.

Após o cálculo dos erros para todos os cenários gerados, uma nova inserção é realizada, agora utilizando todos os dados disponíveis (conjunto original), gerando tantos cenários quantos foram gerados no passo anterior. Assim, o erro máximo relativo obtido em cada cenário será adicionado aos valores complementados proporcionalmente no cenário correspondente. Por exemplo, caso o valor máximo relativo dos erros do cenário 1 seja 0,35, aos valores complementados no banco de dados original no cenário 1 será adicionado 35% do seu valor. Portanto, $z_1^*(u_i) = 22{,}4$ ppm, o novo valor após a transformação fixa será $z_{1'}^*(u_i) = 22{,}4 + (22{,}4 * 0{,}35)$ ppm, caso o viés seja gerado por ausência de valores altos. Assim o valor final MNAR será 30,24 ppm. Se gerados por ausência de valores de baixo teor, o valor corrigido será 14,56 ppm. A figura 4 apresenta um esquema do processo de inserção proposto nesta tese.

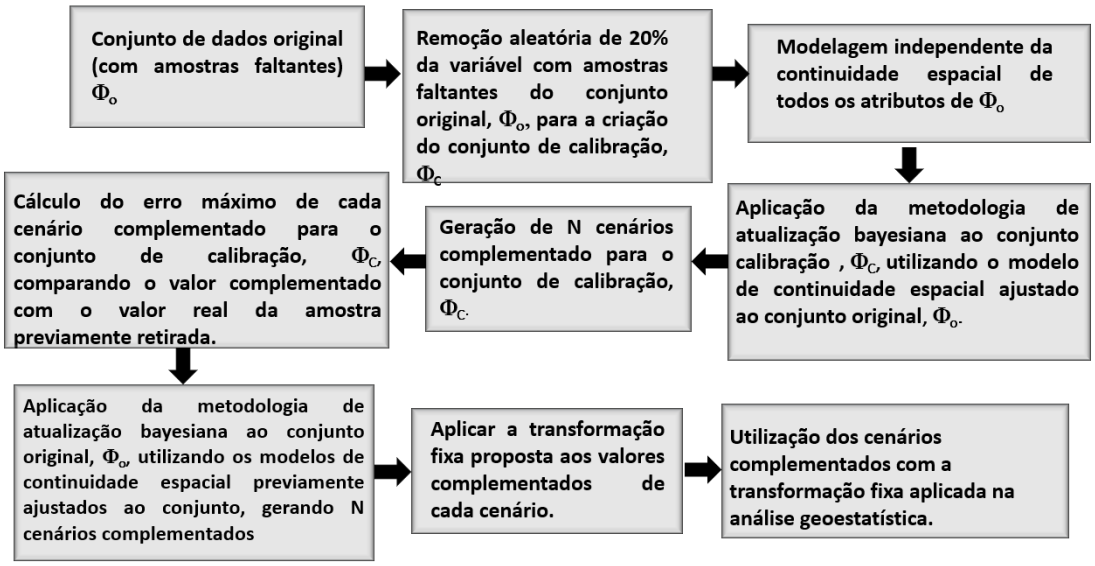

Figura 4 – Esquema de aplicação da metodologia proposta para tratamento de dados faltantes não aleatórios

3.2 Exemplo de aplicação da metodologia de Atualização Bayesiana combinada às transformações fixas.

No exemplo a ser apresentado, foi utilizado o conjunto de dados bidimensional, Jura (Goovaerts, 1997), coletado pelo instituto tecnológico federal em Laussane na Suíça. O conjunto de dados possui 259 amostras, com espaçamento médio de 250 m nas direções norte e leste, representando a concentração em ppm de sete metais: Cádmio (Cd), Cobalto (Co), Cromo (Cr), Cobre (Cu), Níquel (Ni), Chumbo (Pb) e Zinco (Zn).

Para exemplificar o uso da metodologia, foram retiradas amostras de 3 variáveis do conjunto Jura, Co (ppm), Cu(ppm) e Zn (ppm), em quatro regiões distintas das distribuições: P10, P30, P70 e P90, assim na situação de falta P10 foram retiradas 10% das amostras na cauda inferior, na situação P30, 30%. Nos casos P90 e P70, foram retiradas 10% e 30% das amostras respectivamente na cauda superior da distribuição. Cada situação de falta foi analisada separadamente. O conjunto completo será utilizado na análise de resultados como referência, a fim de avaliar o desempenho de cada metodologia aplicada.

A estatística do conjunto completo, para as três variáveis analisadas no conjunto Jura em todos os cenários criados de falta, aqui chamados de casos P10, P30, P70 e P90, é apresentada nas tabelas 2 a 4.

Tabela 2 - Estatística da variável Co (ppm) nos casos completo, P10, P30, P70 e P90.

Variável	Número de amostras	Variância	CV	Média	Assimetria
Co (ppm) completo	259	12,78	0,38	9,30	-0,17
Co (ppm) P10	234	9,80	0,31	9,94	-0,12
Co (ppm) P30	182	4,88	0,19	11,22	0,38
Co (ppm) P90	233	10,17	0,36	8,67	-0,37
Co (ppm) P70	182	7,69	0,36	7,60	-0,29

É possível perceber, a partir da tabela 2, que a variável Co (ppm) possui uma distribuição quase simétrica, uma vez que seu coeficiente de assimetria é negativo, próximo de zero. Além disso, não há indícios de valores extremos na população de dados, e possui coeficiente de variação (CV) no valor de 0,38 no caso completo. Quando são removidas amostras da cauda inferior, o impacto significativo sobre o coeficiente de variação se dá no caso de falta P30, de modo que há uma redução neste valor em torno de 50% e a assimetria da distribuição, ainda que não significativa, passa a ser positiva. Portanto, conclui-se que a ausência de informação neste caso distorce significativamente a distribuição dos dados.

Quando são removidos dados da cauda superior da distribuição, casos P90 e P70, o coeficiente de variação não sofre alteração significativa, sendo que reduz seu valor somente em 5%.

Tabela 3 - Estatística da variável Cu (ppm) nos casos completo, P10, P30, P70 e P90.

Variável	Número de amostras	Variância	CV	Média	Assimetria
Cu (ppm) completo	259	429,013	0,87	23,72	2,84
Cu (ppm) P10	235	436,56	0,81	25,55	2,85
Cu (ppm) P30	182	469,99	0,71	30,14	2,77
Cu (ppm) P90	234	106,34	0,69	18,46	0,98
Cu (ppm) P70	181	30,48	0,39	13,92	0,14

Na tabela 3, nota-se que diferentemente da variável Co (ppm), a variável Cu (ppm) é significativamente assimétrica positiva, em que seu valor de assimetria é 2,84 e possui CV de 0,84 no caso completo. Isto pode indicar a presença de valores extremos na distribuição de probabilidades. É verificado, que a falta exerce mais distorções na distribuição nos casos de P90 e P70, desta forma a distribuição no caso P70 é quase simétrica, com coeficiente de 0,14. Além disso, o valor de CV no caso P90 tem uma redução de 20%, similar ao impacto gerado

quando são removidos 30% dos dados na cauda inferior da mesma distribuição (P30). A distorção observada quando são removidos os dados do P90 e P70 tende a ser mais evidente, uma vez que nesta região estão os valores mais discrepantes da distribuição.

A tabela 4 apresenta a estatística para a terceira variável, Zn (ppm).

Tabela 4 - Estatística da variável Zn (ppm) nos casos completo, P10, P30, P70 e P90.

Variável	Número de amostras	Variância	CV	Média	Assimetria
Zn (ppm) completo	259	842,11	0,38	75,07	1,022
Zn (ppm) P10	234	730,29	0,34	79,46	1,30
Zn (ppm) P30	182	610,56	0,28	87,90	1,65
Zn (ppm) P90	233	419,10	0,30	68,23	-0,004
Zn (ppm) P70	181	250,69	0,26	60,44	-0,32

Na tabela 4, é visto que a variável Zn (ppm), assim como Cu (ppm), possui assimetria positiva, porém o CV, assim como o de Co (ppm), é 0,38. Nos casos em que são retirados valores da distribuição no P30, a distribuição torna-se mais homogênea, o CV passa ao valor de 0,28, tendo uma redução de 26%, contudo a assimetria permanece positiva. Quando são retirados valores de P90 e P70 o comportamento de CV é similar, tendo uma redução no caso de P70 próximo a 26%, todavia a assimetria da distribuição neste caso não é significativa, de modo que há indícios que existem valores extremos na região em que foram retirados os dados.

Nos três casos são observadas distorções significativas quando existem amostras faltantes no conjunto de dados. Diante dessa conformidade é visto que as populações incompletas não são representativas do fenômeno. As correlações existentes entre as variáveis do conjunto também são afetadas pela falta de dados. A tabela 5 apresenta as correlações de Co (ppm) com as variáveis restantes do conjunto de dados nos casos completo e faltantes.

Tabela 5 – Correlações da variável Co (ppm) nos casos completo, P10, P30, P70 e P90.

	Cd	Cr	Cu	Ni	Pb	Zn
Co	0,25	0,45	0,21	0,75	0,18	0,47
Co P10	0,14	0,28	0,17	0,66	0,15	0,35
Co P30	-0,14	-0.13	0,15	0,37	0,031	0,012
Co P90	0,31	0,51	0,22	0,77	0,22	0,53
Co P70	0,39	0,56	0,12	0,78	0,15	0,18

É possível notar a partir da tabela 5, que as correlações entre variáveis são deterioradas devido à falta de dados, no caso completo correlação mais alta que Co (ppm) possui é com a variável Ni (ppm) (ρ=0,75), porém esta correlação quando são removidas amostras nos casos P10 e P30, reduz o grau em 12% e 50% respectivamente. No caso P30 especificamente, a variável já não possui correlação significativa com nenhuma das restantes do conjunto. Nos casos de remoção dos valores no P90 e P70, Co (ppm) mantém a forte correlação com o atributo Ni (ppm). A tabela 6 apresenta os valores de correlações para a variável Cu (ppm).

Tabela 6 - Correlações da variável Cu (ppm) nos casos completo, P10, P30, P70 e P90.

	Cd	Cr	Co	Ni	Pb	Zn
Cu	0,11	0,21	0,21	0,22	0,77	0,57
Cu P10	0,07	0,18	0,18	0,20	0,77	0,53
Cu P30	0,09	0,14	0,12	0,17	0,77	0,49
Cu P90	0,12	0,18	0,34	0,25	0,67	0,51
Cu P70	0,06	0,11	0,19	0,17	0,34	0,39

A tabela 6 mostra que no caso completo, Cu (ppm) possui somente correlação significativa com as variáveis Pb (ppm) e Zn (ppm), sendo a relação com Pb (ppm) mais significativa. No caso em que são removidos dados no P10, P30 e P90, a correlação com Pb (ppm) permanece sendo mais expressiva, contudo quando são retirados dados no caso P70, a relação mais forte passa a ser com Zn (ppm), de maneira que no processo de inserção será considerada somente uma variável correlata, que possui o grau de correlação mais alto em cada caso. A tabela 7apresenta os mesmos dados, para a variável Zn (ppm).

Tabela 7 - Correlações da variável Zn (ppm) nos casos completo, P10, P30, P70 e P90.

	Cd	Cr	Cu	Ni	Pb	Co
Zn	0,66	0,67	0,57	0,63	0,59	0,47
Zn P10	0,65	0,59	0,54	0,53	0,55	0,32
Zn P30	0,56	0,44	0,50	0,30	0,54	-0.001
Zn P90	0,57	0,63	0,46	0,71	0,49	0,61
Zn P70	0,49	0,65	0,39	0,74	0,41	0,68

Como ocorrem para as outras variáveis, são observadas distorções nos graus de correlação que a variável Zn (ppm) possui com as restantes nos casos faltantes, do mesmo modo, a cada caso de inserção, será mantida somente a correlata com maior grau de relação no respectivo caso analisado.

Serão realizados 12 processos de inserção separadamente. Para cada variável faltante, serão complementados valores para os casos P10, P30, P90 e P70 por meio de atualização bayesiana. A inserção de cada variável faltante será realizada utilizando somente uma correlata, com os modelos variográficos ajustados a cada respectivo caso, empregando, porém, a mesma estratégia de busca amostral no procedimento de inserção, de maneira que os modelos possam ser afetados somente devido às amostras faltantes e não aos outros fatores do processo de inferência.

Os pares formados para o caso de faltantes P10 são: Co (ppm) e Ni (ppm); Cu (ppm) e Pb (ppm); Zn (ppm) e Cd (ppm).

No caso P30, são: Co (ppm) e Ni (ppm); Cu (ppm) e Pb (ppm); Zn (ppm) e Cd (ppm).

No caso P90, foram utilizados os seguintes pares: Co (ppm) e Ni (ppm); Cu (ppm) e Pb, Zn (ppm) e Ni (ppm).

No caso de falta P70, foram utilizados os pares: Co (ppm) e Ni (ppm); Cu (ppm) e Zn (ppm) e Zn (ppm) e Ni (ppm).

A metodologia de atualização bayesiana assume que os atributos analisados possuem distribuição gaussiana normal, não obstante variáveis geológicas não possuem esta característica, por esta razão as propriedades foram normalizadas por meio da transformação *nscore* (Bliss, 1934; Verly, 1984; Deustch e Journel, 1998). A transformação foi efetuada por meio do *script* executável do tipo Gslib **nscoremv.exe** (Barnett, 2011). Após os pares foram variografados e seus respectivos modelos utilizados na metodologia de atualização bayesiana (Ren, 2007). Os modelos variográficos utilizados são apresentados nas equações 64 a 87.

Cada variograma foi ajustado utilizando 10 espaçamentos de 250m, com tolerância de espaçamento de 100 metros. Os variogramas experimentais são omnidirecionais, uma vez que os pares são modelados independentemente, e a estratégia de busca amostral será isotrópica e possuem a continuidade espacial exposta nas equações de 64 a 87.

Variogramas dos pares no caso de falta P10:

- Par Co (ppm) e Ni (ppm)

$$\gamma_{Co10} = 0{,}4 + 0{,}6Sph\left(\frac{1{,}5km}{omni}\right) \qquad (64)$$

$$\gamma_{Ni} = 0,2 + 0,8Sph\left(\frac{1,0km}{omni}\right) \qquad (65)$$

- Par Cu (ppm) e Pb (ppm)

$$\gamma_{Cu10} = 0,5 + 0,5Sph\left(\frac{0,35km}{omni}\right) \qquad (66)$$

$$\gamma_{Pb} = 0,5 + 0,3Sph\left(\frac{0,3km}{omni}\right) + 0,2Sph\left(\frac{1,5km}{omni}\right) \qquad (67)$$

- Par Zn (ppm) e Cd (ppm)

$$\gamma_{Zn10} = 0,4 + 0,4Sph\left(\frac{0,3km}{omni}\right) + 0,2Sph\left(\frac{1,7km}{omni}\right) \qquad (68)$$

$$\gamma_{Cd} = 0,4 + 0,2Sph\left(\frac{0,3km}{omni}\right) + 0,4Sph\left(\frac{1,0km}{omni}\right) \qquad (69)$$

Variogramas dos pares no caso de falta P30.

- Par Co (ppm) e Ni (ppm):

$$\gamma_{Co30} = 0,4 + 0,6Sph\left(\frac{0,59km}{omni}\right) \qquad (70)$$

$$\gamma_{Ni} = 0,2 + 0,8Sph\left(\frac{1,0km}{omni}\right) \qquad (71)$$

- Par Cu (ppm) a Pb (ppm)

$$\gamma_{Cu30} = 0,5 + 0,5\left(\frac{0,35km}{omni}\right) \qquad (72)$$

$$\gamma_{Pb} = 0,5 + 0,3Sph\left(\frac{0,3km}{omni}\right) + 0,2Sph\left(\frac{1,5km}{omni}\right) \qquad (73)$$

- Par Zn (ppm) e Cd (ppm)

$$\gamma_{Zn30} = 0,4 + 0,4Sph\left(\frac{0,5km}{omni}\right) + 0,2Sph\left(\frac{1,7km}{omni}\right) \qquad (74)$$

$$\gamma_{Cd} = 0,4 + 0,2Sph\left(\frac{0,3km}{omni}\right) + 0,4Sph\left(\frac{1,0km}{omni}\right) \qquad (75)$$

Variogramas ajustados nos casos de falta P90.

- Par Co (ppm) e Ni (ppm)

$$\gamma_{Co90} = 0,4 + 0,6Sph\left(\frac{1,5km}{omni}\right) \quad (76)$$

$$\gamma_{Ni} = 0,2 + 0,8Sph\left(\frac{1,0km}{omni}\right) \quad (77)$$

- Par Cu (ppm) e Pb (ppm)

$$\gamma_{Cu90} = 0,5 + 0,5Sph\left(\frac{0,6km}{omni}\right) \quad (78)$$

$$\gamma_{Pb} = 0,5 + 0,3Sph\left(\frac{0,3km}{omni}\right) + 0,2Sph\left(\frac{1,5km}{omni}\right) \quad (79)$$

- Par Zn (ppm) e Ni (ppm)

$$\gamma_{Zn90} = 0,2 + 0,4Sph\left(\frac{0,5km}{omni}\right) + 0,4Sph\left(\frac{1,0km}{omni}\right) \quad (80)$$

$$\gamma_{Ni} = 0,2 + 0,8Sph\left(\frac{1,0km}{omni}\right) \quad (81)$$

Variogramas ajustados no caso de falta P70

- Par Co (ppm) e Ni (ppm)

$$\gamma_{Co70} = 0,4 + 0,6Sph\left(\frac{1,1km}{omni}\right) \quad (82)$$

$$\gamma_{Ni} = 0,2 + 0,8Sph\left(\frac{1,0km}{omni}\right) \quad (83)$$

- Par Cu (ppm) e Zn (ppm)

$$\gamma_{Cu70} = 0,55 + 0,45Sph\left(\frac{0,6km}{omni}\right) \quad (84)$$

$$\gamma_{Zn} = 0,4 + 0,4Sph\left(\frac{0,57km}{omni}\right) + 0,2Sph\left(\frac{1,1km}{omni}\right) \quad (85)$$

- Par Zn (ppm) e Ni (ppm)

$$\gamma_{Zn70} = 0,2 + 0,4Sph\left(\frac{0,5km}{omni}\right) + 0,4Sph\left(\frac{1,0km}{omni}\right) \quad (86)$$

$$\gamma_{Ni} = 0,2 + 0,8Sph\left(\frac{1,0km}{omni}\right) \quad (87)$$

Uma vez que os variogramas das variáveis normalizadas são ajustados, é possível utilizar a rotina do estilo Gslib (Deutsch e Journel, 1998) **impute.exe** (Barnett, 2013) por meio da qual a metodologia de atualização bayesiana descrita no capitulo 3 desta tese é aplicada.

A estratégia de busca para cada inserção realizada foi mantida igual, a fim de avaliar o impacto somente das amostras faltantes em variáveis com diferentes graus de CV. Assim, cada variável foi imputada utilizando um máximo de 48 nós previamente simulados, a busca por amostras vizinhas foi realizada de forma isotrópica em um raio de 600 metros. Cada inserção resultou em 30 realizações de cenários completos.

3.2.1 Inserções de Co (ppm), Cu (ppm) Zn (ppm) nos casos P10, P30, P90 e P70

A fim de avaliar a qualidade do modelo aplicado ao conjunto de dados, são retiradas em cada caso de falta, P10, P30, P90 e P70, para cada variável vinte por cento das amostras presentes aleatoriamente. Assim, é possível comparar o valor complementado a cada uma das amostras retiradas ao valor real existente no conjunto. Este grupo de amostras retiradas para teste são aqui referidas como conjunto de calibração. As amostras para teste são retiradas aletoriamente, pois se acaso fossem retiradas de alguma região específica da distribuição poderiam imprimir mais distorções na população. Assim, a retirada aleatória não interfere na estatística da população conhecida.

Aos conjuntos de calibração, também são feitas inserções por atualização bayesiana, utilizando os modelos ajustados para cada par, expressos nas equações 64 a 87, empregando mesma estratégia de busca, bem como o mesmo número de nós previamente simulados e número de realizações de cenários completos. A tabela 8 apresenta o comparativo entre a estatística do conjunto completo com os complementados. Os devios em relação à media e desvio padrão são calculados a partir das definições $Err_m\% = \frac{m_{completo} - m_{complementado}}{m_{completo}}$ e $Err_\sigma\% = \frac{\sigma_{completo} - \sigma_{complementado}}{\sigma_{completo}}$

Tabela 8 – Comparação das estatísticas nos conjuntos de calibração por meio de atualização bayesiana casos P10 e P30 com o caso completo.

Variável	Média (m)	Desvio padrão (σ)	Err_m %	Err_σ %
Co (ppm) conjunto de calibração P10	9,16	3,43	----------	---------
Co (ppm) complementado P10	9,58	3,20	4,8	6,7
Co (ppm) conjunto de calibração P30	11,27	2,21	----------	---------
Co (ppm) complementado P30	11,64	2,37	3,28	7,2
Cu (ppm) conjunto de calibração P10	22,68	14,66	----------	---------
Cu (ppm) complementadoP10	25,12	20,23	-10	-37
Cu (ppm) conjunto de calibração P30	31,81	21,77	----------	---------
Cu complementado (ppm) P30	36,91	26,41	-16	-21
Zn (ppm) conjunto de calibração P10	82,2	28,8		
Zn complementado (ppm) P10	83,58	28,28	-1,6	1,8
Zn (ppm) conjunto de calibração P30	85,35	16,95		
Zn complementado (ppm) P30	87,3	23,90	-2,28	-41

A tabela 8 apresenta os parametros da distribuição das amostras contidas no conjunto de calibração, bem como os parâmetros obtidos para a distribuições obtidas por meio de inserção nos casos testes. Assim, é possível verificar o erro de estimativa de tais parâmetros, antes de proceder com a inserção nos casos faltantes P10 e P30. Com base na tabela 8, vê-se que a estimativa da média possui erros menos significativos que a estimativa do desvio padrão das distribuições. Para Cu (ppm) por exemplo, o erro da estimativa do desvio padrão é de 37% no caso P10. Ainda, no caso de Zn (ppm) com falta em P30, a estimativa do desvio padrão é não satisfatoria, com erro em torno de 41%. Os erros aqui presentes são provenientes de diversos fatores. Cu (ppm) por exemplo, além de possuir amostras faltantes no conjunto, no caso P10 permanece com um valor de CV expressivo, o que pode impactar na qualidade dos valores simulados, porém existe um viés presente nos valores complementados devido ao fato de que o mecanismo de falta primordial do conjunto é MNAR e a metodologia aplicada, atualização bayesiana, é adequada para o mecanismo de falta MAR, podendo conter viéses nos cenários gerados caso a o mecanismo MAR não seja o atuante. A tabela 9 apresenta os resultados obtidos nos casos P90 e P70.

Tabela 9 - Parâmetros estimados nos conjuntos de calibração por meio de atualização bayesiana, casos P90 e P70.

Variável	Média (m)	Desvio padrão (σ)	Err_m %	Err_σ %
Co (ppm) conjunto de calibração P90	9,0	3,20	----------	---------
Co (ppm) complementado 90	8,96	3,15	0,4	1,5
Co (ppm) conjunto de calibração P70	7,70	2,98	---------	---------
Co (ppm) complementado P70	7,3	2,64	5	11
Cu (ppm) conjunto de calibração P90	20,60	11,82	---------	---------
Cu (ppm) complementado P90	18,20	10,09	11	14
Cu (ppm) conjunto de calibração P70	12,98	6,35	---------	---------
Cu complementado (ppm) P70	12,74	6,0	1,8	5,5
Zn (ppm) conjunto de calibração P90	67,30	19,94	---------	---------
Zn complementado (ppm) P90	67,82	19,53	-0,7	2
Zn (ppm) conjunto calibração P70	59,09	14,15	---------	---------
Zn complementado (ppm) P70	60,30	16,37	2	-15

Como nos casos P10 e P30, a tabela 9 apresenta, para os casos P90 e P70, o erro obtido nos parâmetros estimados dos conjutos de calibração. Vê-se, que para a variável Cu (ppm), no caso P90 e para a variável Zn (ppm) no caso P70 houve erro significativo na estimativa dos parâmetros, 11% na estimativa da média e 14% na estimativa do desvio padrão da variável Cu (ppm) e 15 % na estimativa do desvio padrão de Zn (ppm).

Após a inserção no conjunto de calibração, foi realizada a inserção das variáveis Co (ppm), Cu (ppm) e Zn (ppm) em cada caso de falta, P10, P30, P70 e P90. Os valores complementados, nesta etapa, passam pela transformação fixa proposta por Rubin (1987). Para fins de comparação dos modelos resultantes, serão aplicadas duas transformações fixas aos valores complementados, a primeira corrigindo os valores em 20% do valor complementado (Rubin, 1987) para cima ou para baixo, de acordo com o caso analisado. O modelo gerado utilizando a transformação com a constante de 20% é referido como Modelo Rubin. A segunda, os valores serão corrigidos baseado no máximo erro relativo obtido em cada cenário complementado do conjunto de calibração. Nesta, também, somando ou subtraindo do valor complementados de acordo com a equação 63 apresentada na seção 4.1 deste capítulo, aqui referido como modelo erro máximo. Já o modelo de inserção em que não são aplicadas as transformações fixas é referido como modelo de complementado puro.

Ao final, foram analisados os méritos de cada modelo após a aplicação das transformações fixas, do modelo complementado puro em relação ao conjunto de dados referência Jura, que possui todas as amostras medidas.

A figura 5 Apresenta o histograma acumulado da variável Co (ppm) contendo as 259 amostras do conjunto referência, Jura, bem como o histograma médio das 30 realizações em cada metodologia de inserção, modelo Rubin, modelo erro máximo e complementado puro.

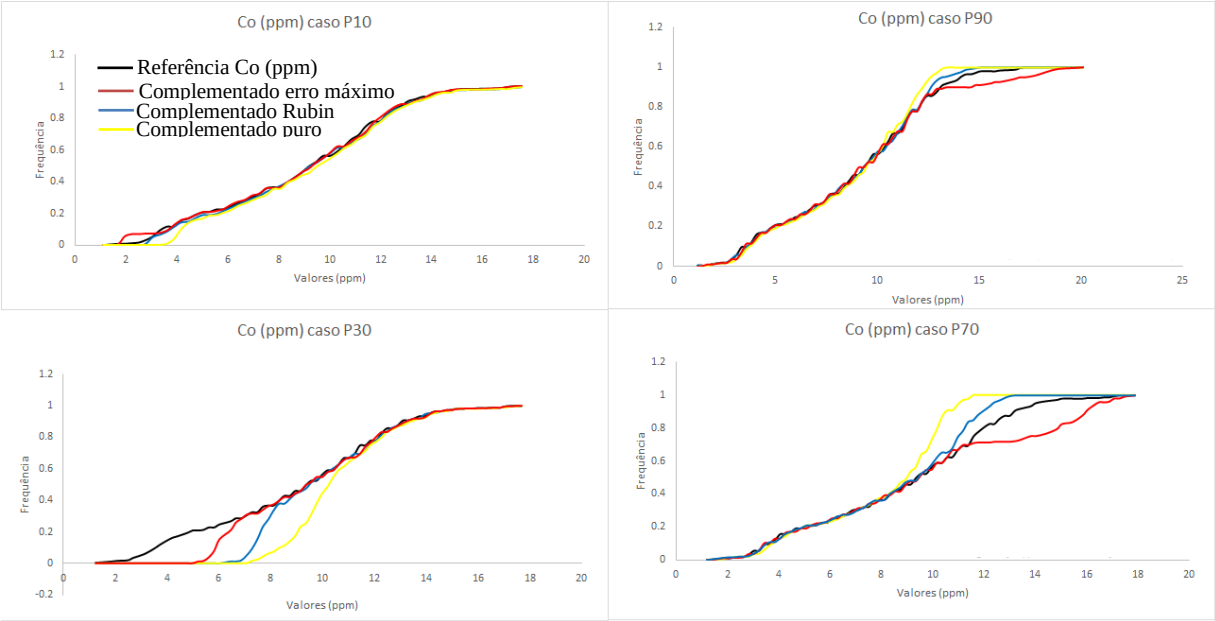

Figura 5 – Histograma acumulado dos dados Co (ppm) referência Jura (curva preta), Modelo Rubin (curva azul), modelo erro máximo (curva vermelha) e inserção pura (curva amarela)

A partir da figura 5, nota-se que no caso P10 os histogramas dos dados completos (referência Jura) é bem reproduzido por todas a metodologias de inserção, já nos casos P30 e P70, onde a falta de dados é 30% nenhuma das metodologias obteve resultados satisfatórios no que se refere à reprodução da curva de histograma dos dados completos, sendo o melhor desempenho observado pelo modelo Rubin.

A tabela 10 apresenta as estimativas médias dos parâmetros da distribuição dos dados obtidas por meio da inserção pura e da inserção aplicadas as transformações fixas de Rubin, utilizando a constante de 20% e proporcional ao erro máximo obtido.

Tabela 10 – Comparativo das estatísticas da variável Co (ppm) estimados pelos modelos complementados *(e-type)* em relação ao conjunto referencia Jura.

	Média (m)	Desvio padrão (σ)	Err_m %	Err_σ %
Co (ppm) referência (Jura)	9,30	3,57	------	------
Co (ppm) P10 Complementado puro	9,48	3,33	-1,9	6,7
Co (ppm) P10 Modelo Rubin	9,38	3,46	-0,8	3
Co (ppm) P10 Modelo erro máximo	9,27	3,63	0,3	-1,6
Co (ppm) P30 Complementado puro	10,84	1,97	-16	44
Co (ppm) P30 Modelo Rubin	10,25	2,39	-10	33
Co (ppm) P30 Modelo erro máximo	9,81	2,86	-5,4	19
Co (ppm) P90 Complementado puro	8,93	3,13	3,9	12
Co (ppm) P90 Modelo Rubin	9,5	3,36	-2,1	5,8
Co (ppm) P90 Modelo erro máximo	9,57	4,06	2,9	13
Co (ppm) P70 Complementado puro	8,25	2,57	11	28
Co (ppm) P70 Modelo Rubin	8,83	3,04	5	14
Co (ppm) P70 Modelo erro máximo	10,21	4,45	9,7	-24

É observado na tabela 10 que a inserção pura gera uma estimativa dos parâmetros da distribuição satisfatória somente no caso P10, nos casos em que são faltantes 30% das amostras tanto na cauda inferior (P30) quanto na cauda superior (P70) os erros relativos nas estimativas dos parâmetros são mais significativos que nos casos em que são aplicadas as transformações fixas aos valores complementados (modelo Rubin e modelo erro máximo). No caso P30, o erro relativo na estimativa da média é de 16% no modelo complementado puro, enquanto no modelo Rubin este erro é reduzido para 10% e no modelo erro máximo para 5,4%. A estimativa do desvio padrão no mesmo caso obteve erro de 44% no modelo complementado puro, enquanto no modelo Rubin é de 33% e no modelo erro máximo é de 19%, representando uma redução no erro de 25% e 57% respectivamente. Já no caso P70, o modelo que possui o menor erro relativo é o modelo Rubin, apresentando um erro de 5% na estimativa da média e de 14% na estimativa do desvio padrão.

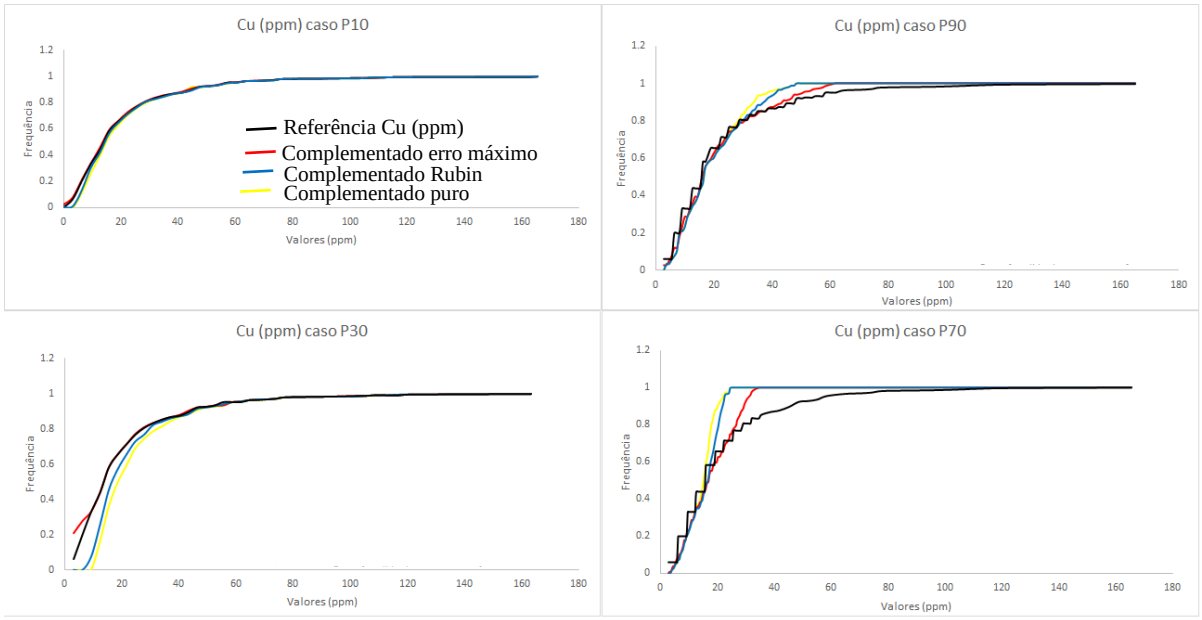

Figura 6 - Histograma acumulado dos dados Cu (ppm) referência Jura (curva preta), Modelo Rubin (curva azul), modelo erro máximo (curva vermelha) e inserção pura (curva amarela)

Assim como no caso da variável Co (ppm), nota-se na figura 6 que no caso de falta P10 os modelos complementados reproduzem de maneira satisfatória a curva de histograma do conjunto completo, porém no caso P90, vemos que o modelo que melhor reproduz a curva de histograma dos dados é o modelo erro máximo, bem como nos casos P30 e P70. É importante ressaltar que a variável Cu (ppm) possui forte assimetria positiva e possui CV de 0,84 no conjunto completo, e como mostrado anteriormente, na análise exploratória dos conjuntos de dados, esta variável sofre forte impacto pela perda de amostras, especialmente nos casos em que a falta se dá na região da cauda superior da distribuição, dificultando o procedimento de inserção dos faltantes.

Tabela 11 - Comparativo das estatísticas da variável Cu (ppm) estimados pelos modelos complementados *(e-type)* em relação ao conjunto referencia Jura.

	Média (m)	Desvio padrão (σ)	Err_m %	Err_σ %
Cu (ppm) referência (Jura)	23,72	20,71		
Cu (ppm) P10 Complementado puro	24,58	20,19	-3,6	2,5
Cu (ppm) P10 Modelo Rubin	24,30	20,32	-2,4	1,8
Cu (ppm) P10 Modelo erro máximo	23,63	20,79	0,37	0,38
Cu (ppm) P30 Complementado puro	27,79	19,01	-17	8,2
Cu (ppm) P30 Modelo Rubin	26,47	19,32	-11	6,7
Cu (ppm) P30 Modelo erro máximo	22,91	21,30	3,4	-2,8
Cu (ppm) P90 Complementado puro	19,62	10,49	-17	49
Cu (ppm) P90 Modelo Rubin	20,20	11,26	14	45
Cu (ppm) P90 Modelo erro máximo	21,73	14,17	8,3	31
Cu (ppm) P70 Complementado puro	14,76	4,88	37	76
Cu (ppm) P70 Modelo Rubin	15,76	5,53	33	73
Cu (ppm) P70 Modelo erro máximo	18,38	8,39	22	59

Na tabela 11, percebe-se que as estimativas das estatísticas para a variável Cu (ppm), o erro relativo é menor em todos os casos para o modelo erro máximo, porém ainda apresentam valores altos de erros relativos em todos modelos. Isto se deve ao fato que não somente existem amostras faltantes no conjunto de dados em uma região específica da distribuição, dificultando portanto a estimativa de uma distribuição representativa dos dados, mas Cu (ppm) é um atributo que possui forte assimetria e valores extremos. Os dados retirados da cauda superior continham os valores que causavam a assimetria da distribuição, gerando distorções que dificilmente serão corrigidas baseadas nas amostras restantes do conjunto.

A figura 7 apresenta os histogramas cumulativos médios dos modelos complementados e o histograma do conjunto completo do atributo Zn (ppm).

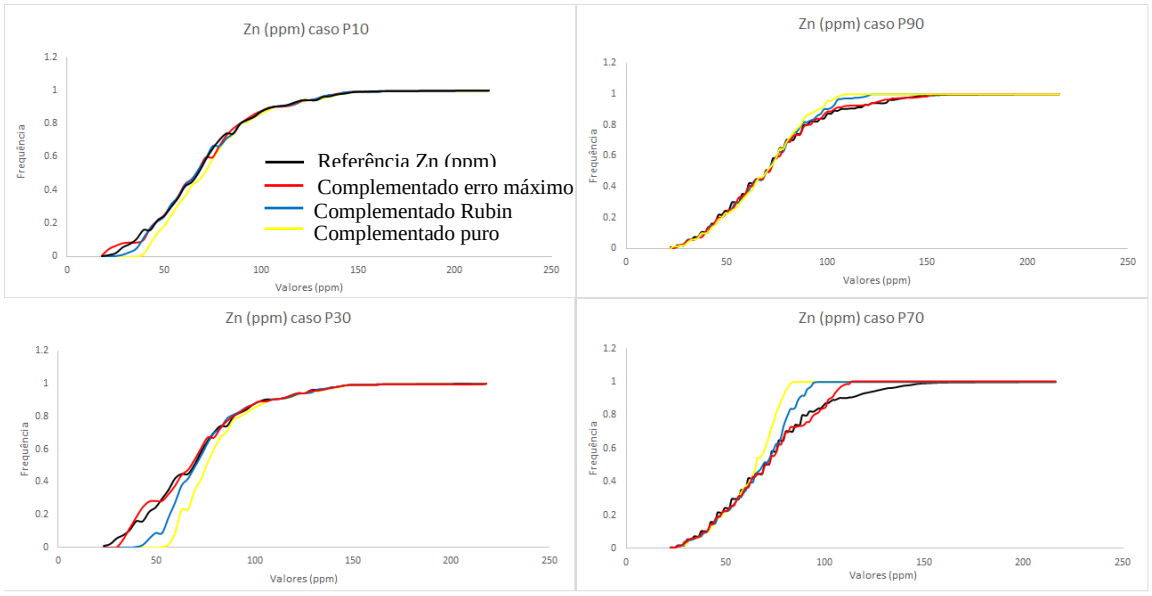

Figura 7 - Histograma acumulado dos dados Zn (ppm) referência Jura (curva preta), Modelo Rubin (curva azul), modelo erro máximo (curva vermelha) e inserção pura (curva amarela)

Na figura 7, nota-se que a curva do histograma acumulado dos dados Zn (ppm) no conjunto referência é honrado em todos os casos de falta corrigido pelo modelo erro máximo, com exceção do caso de falta P70, onde ainda apresenta melhor ajuste que os modelos Rubin e complementado puro. A tabela 12 apresenta um sumário da estimativa dos parâmetros média e desvio padrão pelo E-type das realizações de cada modelo complementado, comparado aos parâmetros do conjunto completo de Zn (ppm).

Tabela 12 - Comparativo das estatísticas da variável Cu (ppm) estimados pelos modelos complementados *(e-type)* em relação ao conjunto referencia Jura.

	Média (ppm) (m)	Desvio padrão(ppm) (σ)	Err_m %	Err_σ %
Zn referência (Jura)	75,07	29,07		
Zn P10 Complementado puro	77,18	26,91	-2,8	7,4
Zn P10 Modelo Rubin	76,10	27,85	-1,2	4,1
Zn P10 Modelo erro máximo	74,72	29,58	0,46	-1,7
Zn P30 Complementado puro	84,13	22,20	-12	23
Zn P30 Modelo Rubin	79,66	24,64	6,1	15
Zn P30 Modelo erro máximo	74,70	29,18	0,4	-0,3
Zn P90 Complementado puro	69,99	20,48	7	29
Zn P90 Modelo Rubin	71,71	22,52	4,4	22
Zn P90 Modelo erro máximo	74,23	27,09	1,1	6,8
Zn P70 Complementado puro	63,60	14,53	15	50
Zn P70 Modelo Rubin	67,88	17,95	9,5	38
Zn P70 Modelo erro máximo	72,54	23,28	3,3	19

A tabela 12 mostra que, neste caso, também se obtiveram melhores estimativas dos parâmetros da distribuição por meio do modelo erro máximo em todos os casos, ainda que a variável Zn (ppm) possua assimetria positiva.

Ainda, foram verificadas a acurácia e precisão dos modelos, complementado puro e os modelos em que foram aplicadas as transformações fixas, modelo Rubin e modelo erro máximo. Considerando a amostra $z_f(u_i)$, como a amostra faltante no local u_i, $z_o(u_i)$ o valor medido no local u_i, e $z_j^*(u_i)$ a j-ésima inserção do valor no local u_i é possível obter uma distribuição de possíveis valores para cada local. Assim, as 30 realizações foram ordenadas e para os intervalos de probabilidade de 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20% e 10% foi verificada a proporção em que os valores reais estão contidos no intervalo de valores realizados. Assim pode-se obter um gráfico de intervalo de probabilidade versus proporção de ocorrência. As figuras 8 a 11 apresentam os gráficos de acuracidade para cada variável em cada caso de falta para os 3 modelos aplicados aos dados, complementado puro, Rubin e erro máximo.

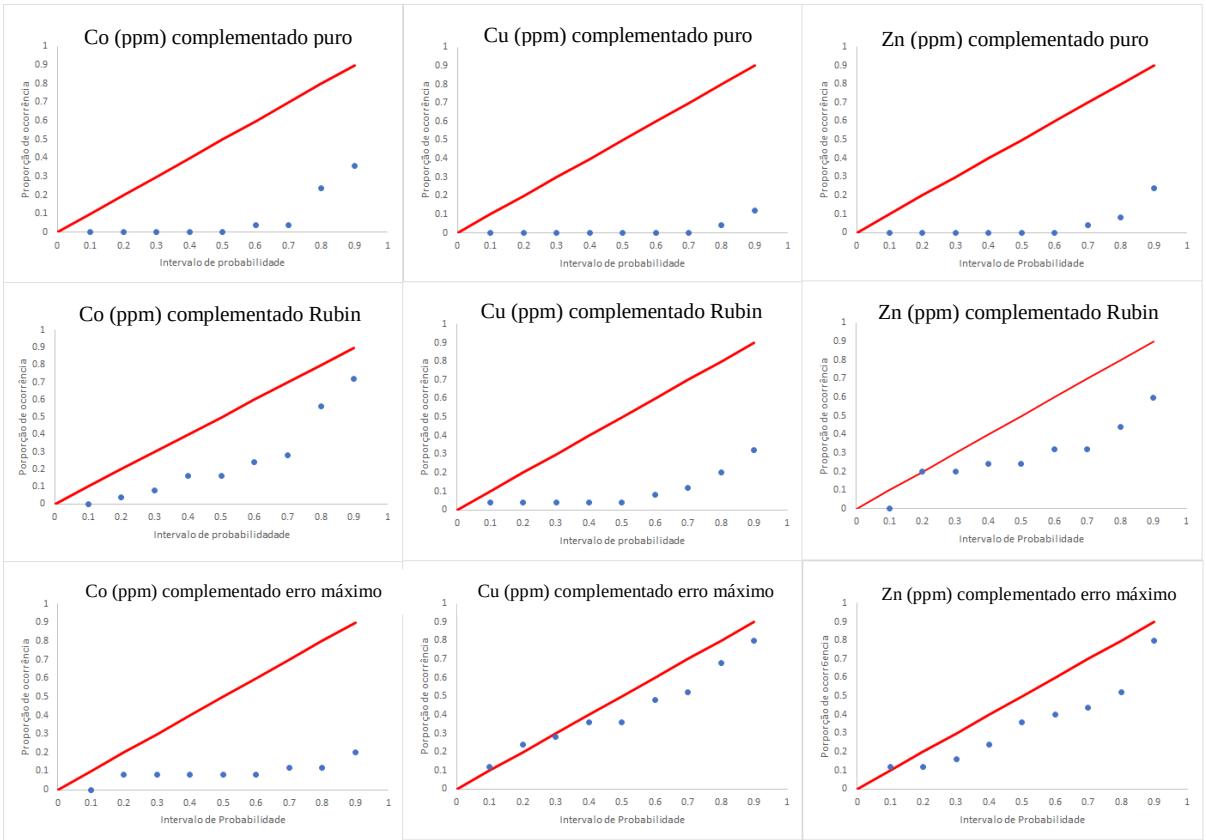

Figura 8– Gráficos de acuracidade no caso de falta P10, atributo Co(ppm), Cu (ppm) e Zn (ppm)

Na figura 8, são apresentados gráficos de acuracidade para as três variáveis no caso de falta P10. É observado que em todos os casos os modelos Rubin e erro máximo, apresentam melhora, ou seja, há maior proporção de ocorrência dos valores reais nos intervalos de probabilidade que os obtidos no modelo complementado puro. No caso de Cu (ppm) por exemplo, é visto que no modelo complementado puro, a proporção de ocorrência do valor real dentro do intervalo de probabilidade 90% é de 10% enquanto nos modelos Rubin e erro máximo é de 33% e 80% respectivamente. Para a variável Co (ppm), a melhora significativa ocorre no Modelo Rubin, onde para o intervalo 90% ocorre uma proporção de 72% de ocorrência. A figura 9 apresenta os gráficos de acuracidade no caso de falta P30.

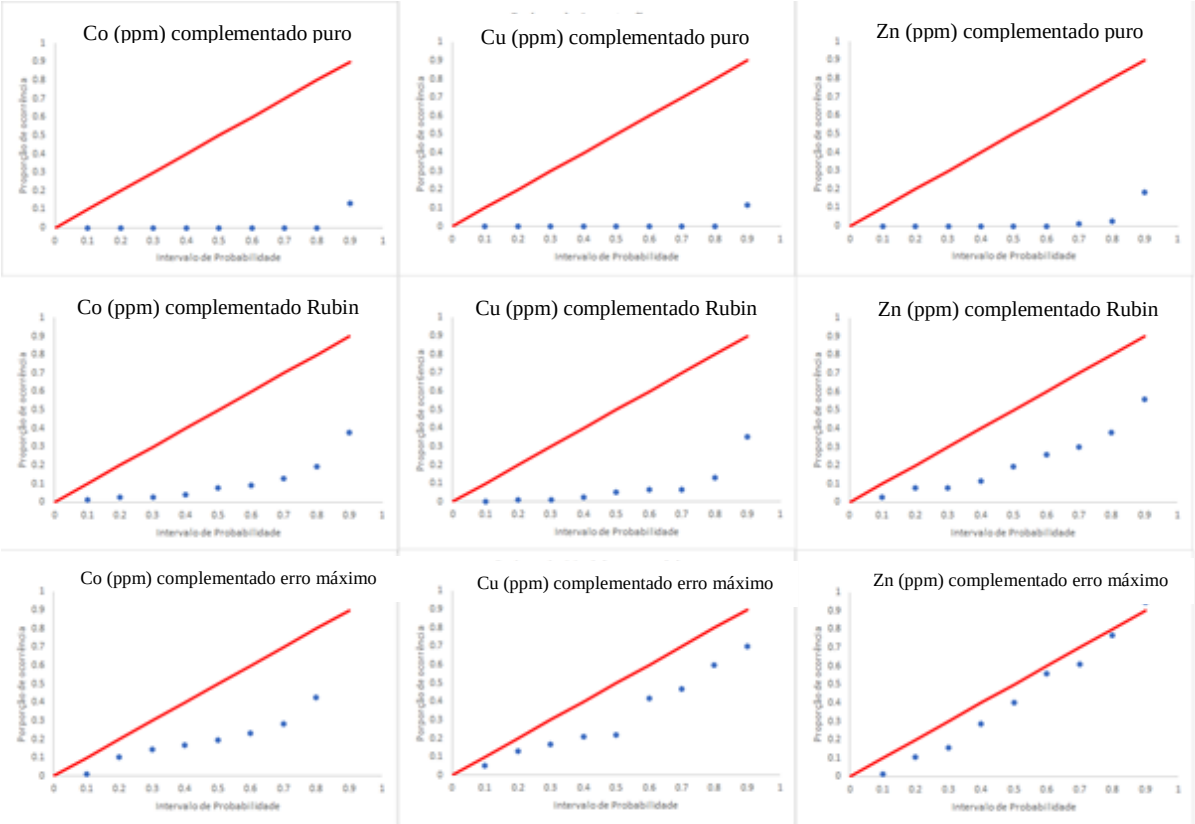

Figura 9- Gráficos de acuracidade no caso de falta P30, atributo Co(ppm), Cu (ppm) e Zn (ppm)

Na figura 9, é observado que o modelo que obteve maiores proporções de ocorrência dentro dos intervalos de probabilidade foi o modelo erro máximo, para todas as variáveis. Contudo, o modelo Rubin oferece progresso em relação ao modelo de complementado puro que possui proporções de ocorrência muito baixas em todos os intervalos de probabilidade. A figura 10 apresenta os gráficos de acuracidade para o caso de falta P90.

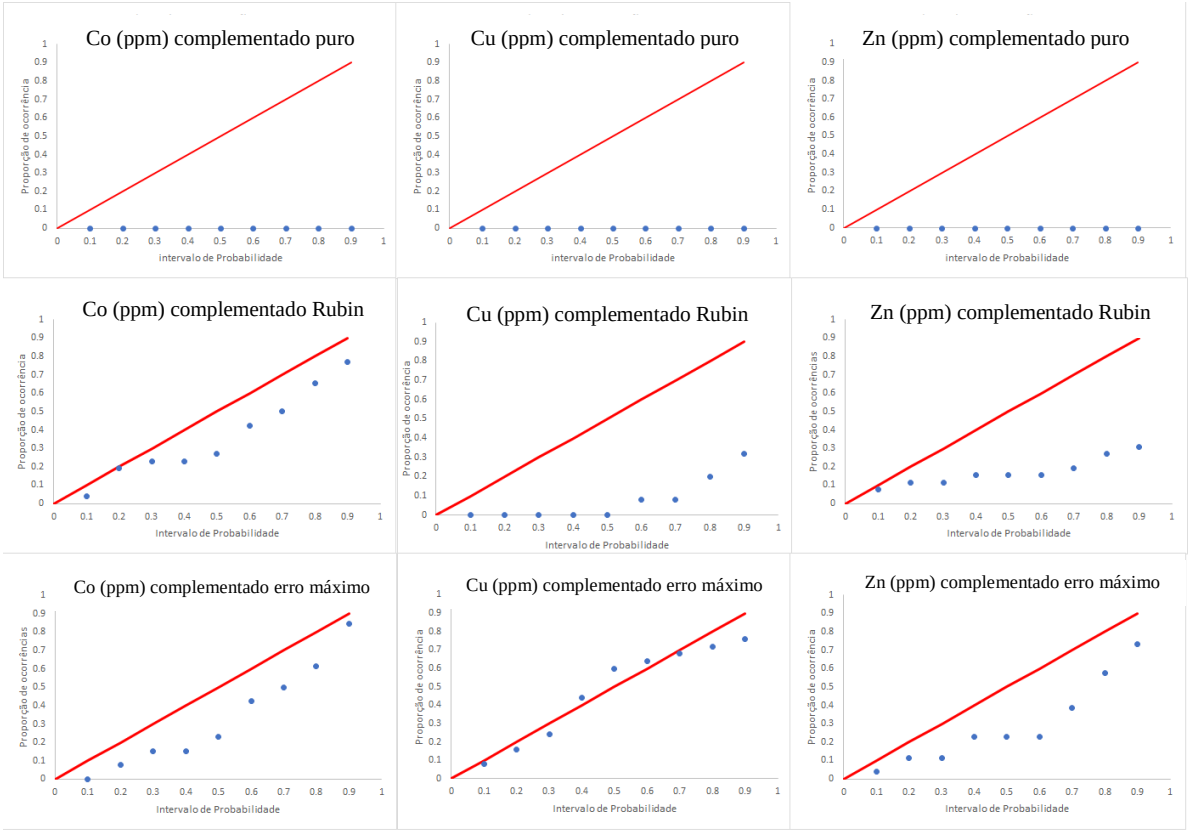

Figura 10 - Gráficos de acuracidade no caso de falta P90, atributo Co(ppm), Cu (ppm) e Zn (ppm)

É observado na figura 10 que no caso de falta P90, assim como no caso P30, o modelo erro máximo apresenta maiores proporções de ocorrência dos valores reais nos intervalos de probabilidade estipulados que nos modelos Rubin e inserção pura para as variáveis Cu (ppm) e Zn (ppm). O modelo Rubin apresenta melhor desempenho para a variável Co (ppm), resultados consistentes com os histogramas acumulados apresentados anteriormente. A figura 11 mostra os gráficos de acuracidade para o caso de falta P70.

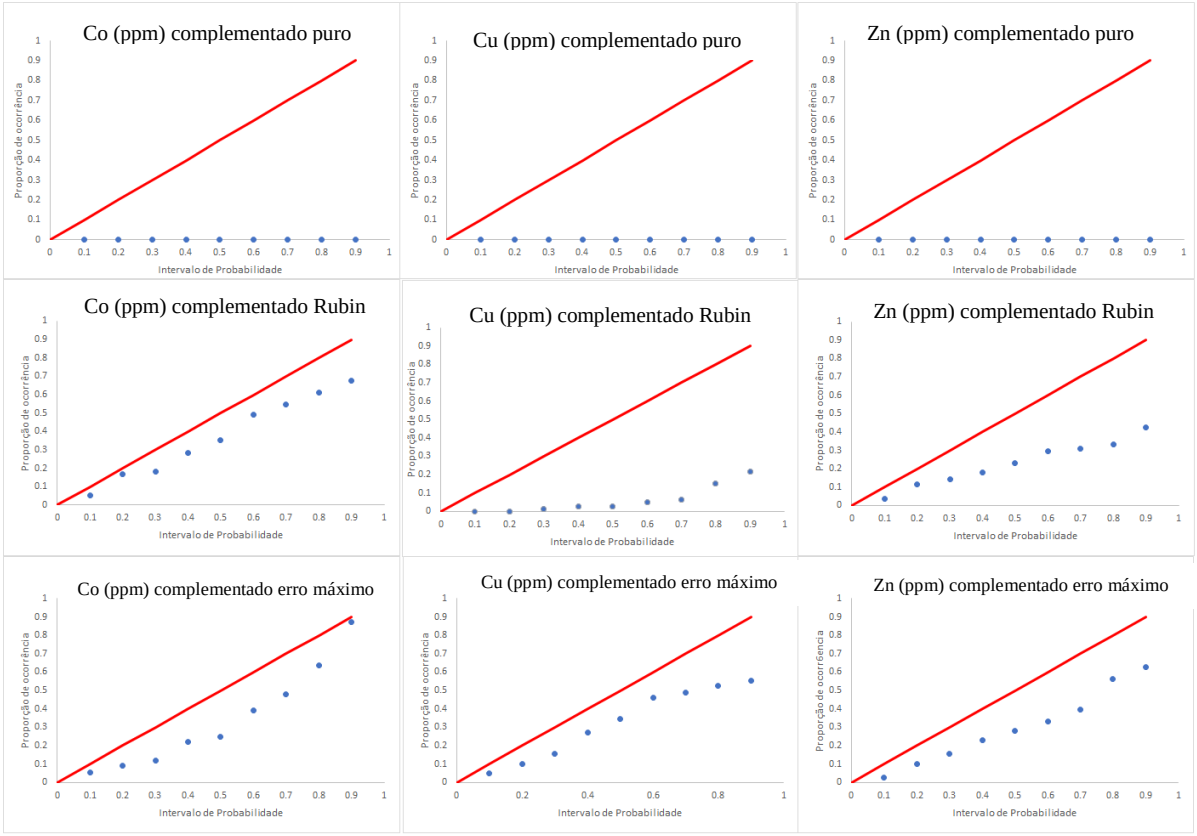

Figura 11 - Gráficos de acuracidade no caso de falta P70, atributo Co (ppm), Cu (ppm) e Zn (ppm)

Nota-se na figura 11, que há ganhos significativos na aplicação de transformações fixas aos valores complementados, uma vez que o modelo complementado puro não possui para nenhuma das variáveis, ocorrências dos valores reais dentro dos intervalos de probabilidade. Nota-se que o modelo é completamente enviesado em relação à população representativa. Já nos modelos Rubin e erro máximo os valores reais ocorrem dentro dos intervalos.

Considerando os resultados observados neste capitulo verifica-se que o uso de transformações fixas aos valores complementados contribui na construção de modelos geoestatísticos mais representativos do fenômeno em análise, quando há amostras faltantes sob mecanismo MNAR no conjunto de dados.

De acordo com os resultados obtidos neste estudo de caso pode-se verificar que acima de 10% de faltantes há distorções significativas nos valores complementados, requerendo a aplicação de transformações fixas. No caso de Co (ppm), os modelos transformados Rubin e erro máximo apresentaram resultados similarmente satisfatórios quando a falta de dados é de 30%, tanto na cauda superior quanto na cauda inferior. No caso P30, há ligeira melhora no modelo onde foi aplicada a transformação fixa do erro máximo, uma vez que a assimetria mesmo que leve, é negativa. No caso P70 o modelo Rubin apresenta melhores estimativas para

a variável Co (ppm), pois a quantidade de dados faltantes é significativa e gera viés nos valores complementados, porém o viés gerado neste caso é menos influente que no caso P30, deste modo a correção por 20% do valor complementado se mostra suficiente.

Já para as variáveis Cu (ppm) e Zn (ppm) o modelo erro máximo apresenta resultados mais satisfatórios que os modelos Rubin e complementado puro, isso se deve ao fato de que ambos atributos possuem assimetria bastante acentuada. Nos casos faltantes P10 e P30, o modelo erro máximo apresenta erros relativos menores que 5%. No caso P90 e P70, o modelo erro máximo também apresenta ganhos em relação aos outros dois modelos, apresentado maior ocorrência de valores dentro dos intervalos de probabilidade para ambas variáveis.

Desta forma, vê-se que além quantidade de dados faltantes, o grau de assimetria e região de falta da distribuição associada exercem influência nos valores complementados pelo modelo de atualização bayesiana, e podem orientar a escolha da transformação fixa aplicada. A figura12, apresenta um esquema que sumariza os comportamentos e respostas observados neste estudo de caso, bem como a transformação fixa recomendada.

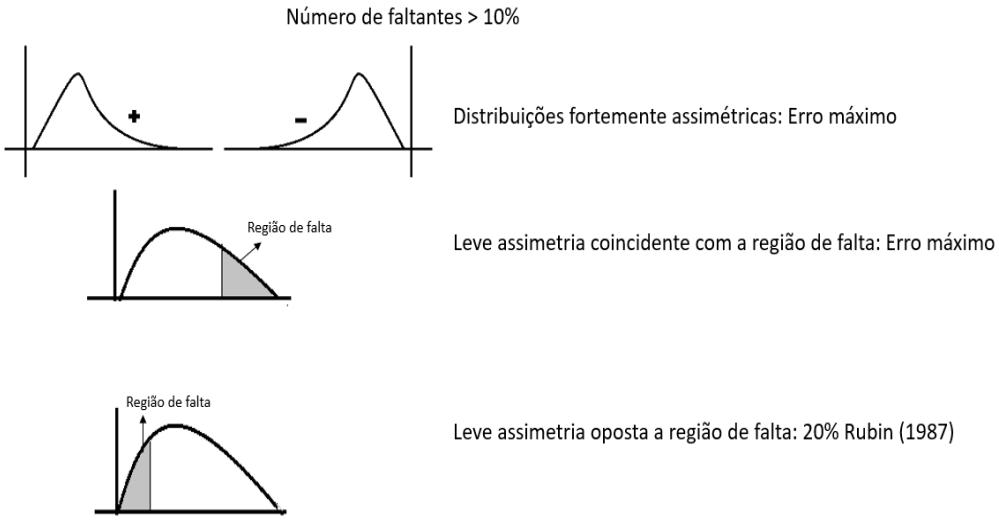

Figura 12 – Esquema de recomendações de uso das transformações fixas, de acordo com estudo de caso apresentado

Como exposto na figura 12, variáveis com assimetria significativa apresentam melhores respostas quando aplicada a transformação fixa do erro máximo, além disso se a região de faltantes for no mesmo sentido que a assimetria, por exemplo, assimetria positiva e falta de dados na região superior da distribuição recomenda-se, baseado neste estudo de caso a aplicação da transformação fixa do erro máximo.

Além disso, baseado nos resultados aqui expostos para variáveis aproximadamente simétricas em que a falta de dados se dá da região oposta à assimetria destas, a aplicação da transformação fixa proposta por Rubin (1987) leva a resultados mais satisfatórios que os outros dois modelos analisados e, portanto, recomenda-se o seu uso.

CAPÍTULO 4

Modelo geoestatístico de variáveis faltantes MNAR

Este capítulo apresenta um estudo de simulação sequencial gaussiana aplicada ao conjunto de dados Walker lake (Isaaks e Srivastava, 1989), onde o conjunto utilizado será complementado por atualização bayesiana e transformado utilizando o erro máximo, como apresentado no capítulo 4 desta tese.

4.1 Modelagem geoestatística conjunto de dados Walker lake

O estudo de caso foi realizado no conjunto de dados amostrais Walker lake (Isaaks e Srivastava, 1989), que possui 470 locais amostrados. Este conjunto de dados foi construído em três etapas, a primeira 195 amostras foram coletadas em uma malha de aproximadamente 20 x 20m da variável V (ppm) e T (categórica). Na segunda etapa amostral uma malha mais densa, em torno de 10 x 10m foi executada nas áreas próximas às amostras de V que haviam sido medidas acima de 500ppm. A terceira etapa de campanha amostral teve por objetivo delinear as zonas de alto valor de V (ppm), coletando adicionalmente 125 amostras na direção leste-oeste o que totalizou 470 amostras das variáveis V (ppm) e T (categórica). A variável U (ppm) foi coletada somente na segunda e terceira etapa de modo que desta faltam 195 amostras.

No presente estudo serão analisadas somente as variáveis contínuas V (ppm) e U (ppm), o mapa amostral para estas duas variáveis é apresentado na figura 13.

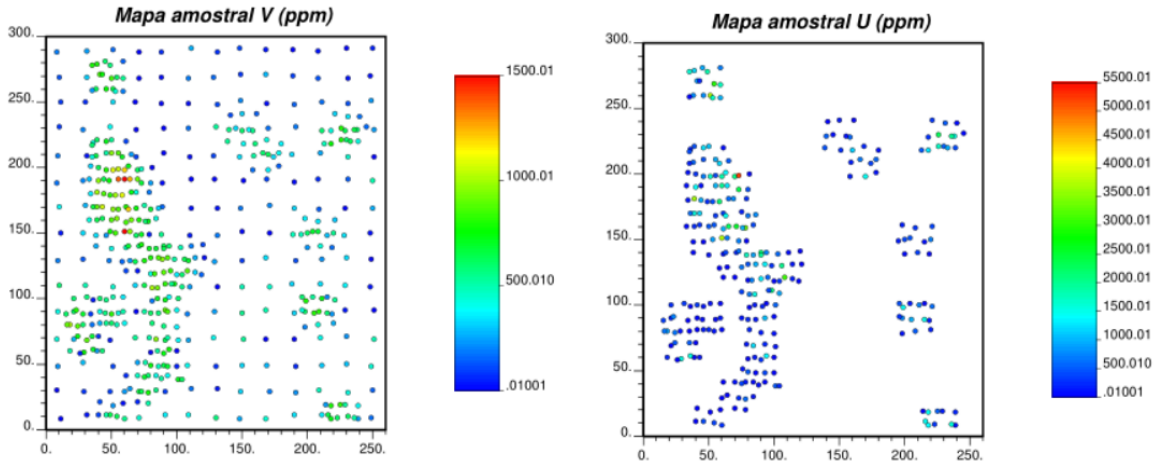

Figura 13– Mapa amostral das variáveis V (ppm) e U (ppm) do conjunto de dados Walker Lake (Isaaks e Srivastava, 1989)

Na figura 13, é verificada a amostragem de U (ppm) em zonas ricas de modo que a malha de 20 x20m não é preenchida por esta variável. Já, V(ppm) possui amostras em toda a área de estudo. Além da ausência de amostras da variável U (ppm), que pode levar a distorções da estatística da distribuição de probabilidade da variável, ambas estão sujeitas ao agrupamento amostral que por sua vez também pode conduzir a estatística não representativa da população. Portanto, é aplicado o método de células móveis a fim de obter uma estatística representativa. Neste caso, foram utilizadas células de 20 x 20m, que corresponde ao espaçamento amostral médio da malha da primeira campanha amostral. A figura 14 apresenta os histogramas agrupados e desagrupados das variáveis V (ppm) e U (ppm).

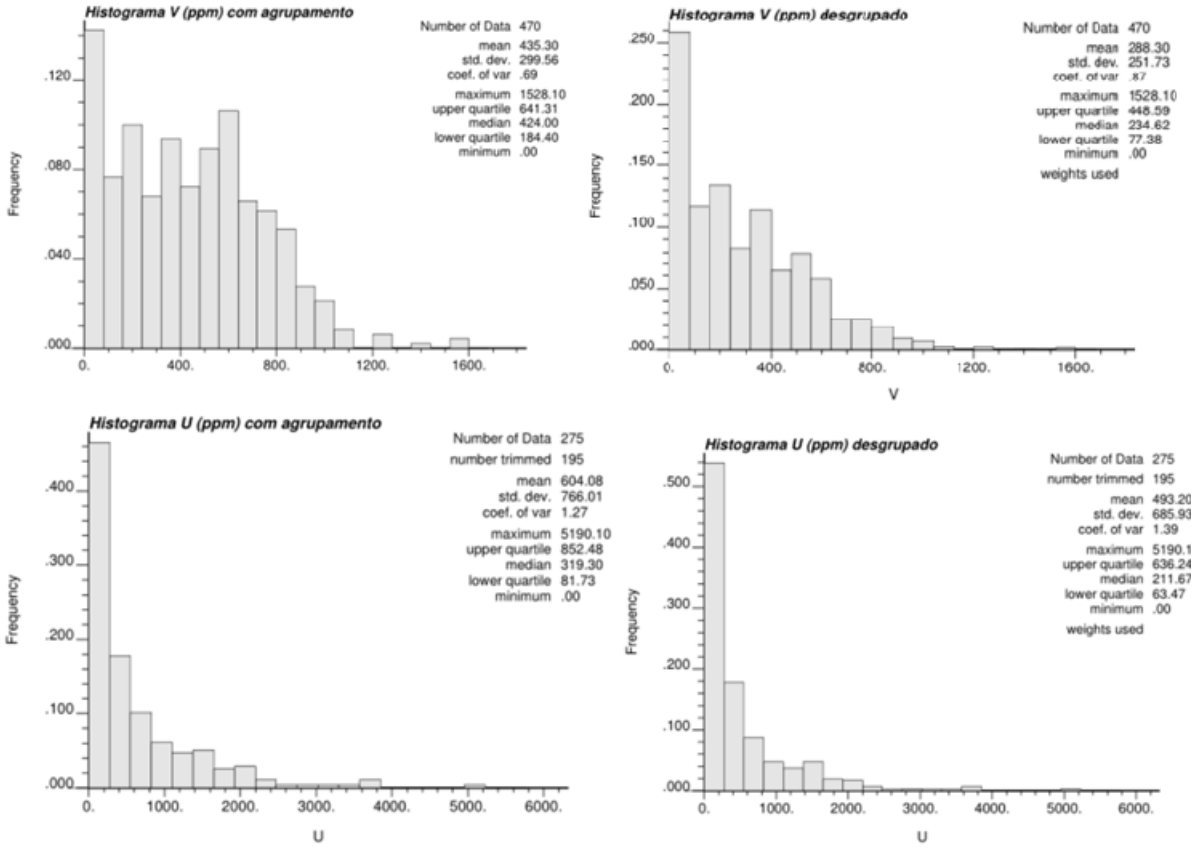

Figura 14 - histogramas agrupados e desagrupados das variáveis V (ppm) e U(ppm) no conjunto de dados Walker Lake

O histograma da variável V (ppm) possui as 470 amostras presentes no conjunto de dados, possui assimetria positiva evidente, e coeficiente de variação alto. A partir da figura 14 é possível notar que o adensamento amostral que ocorreu na segunda e terceira campanha leva a um valor de média superestimado para a população em análise, 435,30 ppm. Após o procedimento de desagrupamento dos dados por meio de células móveis, a média da variável

V (ppm) reduz para 288,30 ppm. A variância também sofre redução após o desagrupamento, enquanto o coeficiente de variação sofre aumento em seu valor.

Assim como para a variável V (ppm), o atributo U (ppm) também possui um valor de média da população superestimado, consistente com a estratégia amostral. Após a aplicação do método de células móveis ao conjunto amostral de U (ppm), a média desta variável reduz de 604,08ppm para 493,20ppm. A variância neste caso também apresenta redução do seu valor original e o coeficiente de variação sofre aumento. O atributo U (ppm) também possui assimetria positiva evidente, e média duas vezes maior que a mediana, indicando a presença de valores extremos no conjunto de dados. Além disso, as duas variáveis possuem correlação positiva no valor de 0,55.

O objetivo deste estudo de caso é realizar a modelagem geoestatística da variável U (ppm), que é sub amostrada no conjunto de dados. Desta maneira, será aplicada a metodologia proposta nesta tese, que consiste em preencher os locais de amostras faltantes por meio de atualização bayesiana e, visto que os dados faltantes pertencem a uma região da distribuição de U(ppm), aplicar a transformação fixa baseada no erro máximo. Além disso o conjunto será modelado também aplicando a transformação fixa proposta por Rubin (1987), vinte por cento do valor complementado, será avaliada a performance do modelo complementado puro e da abordagem clássica de cossimulação das variáveis. O conjunto Walker lake exaustivo (Isaaks e Srivastava, 1989) será utilizado para fins de validação dos modelos construídos por meio de cada metodologia.

4.1.1 Atualização bayesiana aplicada ao conjunto Walker Lake

Como visto no capítulo 2, a metodologia de atualização bayesiana assume a gaussianidade do conjunto de dados, portanto requer, inicialmente, a transformação *nscore* (Bliss, 1934; Verly, 1984; Deustch e Journel, 1998). É aplicada às variáveis por meio da rotina estilo Gslib (Deutsch e Journel, 1998), **nscoremv.exe** (Barnett, 2011). Uma vez que os dados possuem sua distribuição normalizada, a continuidade espacial é modelada. É importante ressaltar que a rotina de atualização bayesiana não requer a modelagem de covariância cruzada, somente os modelos diretos.

Cada modelo variográficos foi construído utilizando 10 espaçamentos, como 20m cada espaçamento e tolerância de 10metros. Os modelos foram ajustados nas direções N157, de

maior continuidade espacial, e na N67, de menor continuidade espacial. A largura de banda utilizada foi de 10m.

Os modelos ajustados aos variogramas experimentais são expressos nas equações 88 e 89.

$$\gamma_U(h) = 0{,}45 + 0{,}3\, Sph\left(\frac{34m}{N157}\frac{16m}{N67}\right) + 0{,}25 Sph\left(\frac{122m}{N157}\frac{34m}{N67}\right) \qquad (88)$$

$$\gamma_V(h) = 0{,}3 + 0{,}3 Sph\left(\frac{40m}{N157}\frac{20m}{N67}\right) + 0{,}4 Sph\left(\frac{114m}{N157}\frac{34m}{N67}\right) \qquad (89)$$

A estratégia de busca de amostras para a realização da atualização bayesiana foi efetuada nas direções preferenciais dos modelos variográficos ajustados, N157 e N67, com raios de busca equivalentes aos alcances da variável sub amostrada, 122m na direção N157 e 34m na direção N67. Foram executadas 50 realizações de cenários completos por meio da atualização bayesiana.

Para o cálculo do erro máximo e aplicação da transformação fixa, é necessário realizar a inserção também a um conjunto de calibração. Este refere-se a retirada de 20% das amostras de U (ppm) do conjunto conhecido Walker Lake, para fins de acesso ao erro dos valores complementados. O modelo de continuidade espacial, exposto nas equações 112 e 113 é aplicado também a este conjunto de calibração. Deste modo, a atualização bayesiana é empregada ao conjunto de calibração.

O procedimento resulta em diversos cenários em que as amostras de calibração foram imputadas. Baseado nos valores complementados destas amostras, e em seus valores conhecidos a partir do conjunto original Walker lake, é possível acessar o erro relativo em cada local e em cada cenário gerado. O valor máximo obtido a cada cenário é adotado como parâmetro na transformação fixa dos valores complementados para a variável U (ppm) do conjunto original Walker lake.

Assim foi criado o modelo complementado onde foi empregada a transformação fixa proposta nesta tese, segundo a equação 87 do capitulo 3.

Além do modelo supracitado, também foi gerado o modelo complementado onde foi aplicada a transformação fixa de proposta por Rubin, alterando em 20% cada valor complementado nos cenários realizados.

Ao fim da análise de atualização bayesiana, resultam três modelos distintos com 50 cenários completos cada, em que foram inseridos valores a cada um dos 195 locais em que se tem presente somente medidas da variável V (ppm): um modelo complementado puro, modelo erro máximo e modelo de Rubin. A estatística do cenário médio de cada modelo é apresentada na tabela 13, onde o cenário médio será utilizado na modelagem geoestatística do conjunto.

Tabela 13 – estatística dos cenários médios dos modelos complementados, puro, Rubin e erro máximo.

	Média (ppm)	**Variância**	**CV**	**Assimetria**	ρ_{UV}
Modelo complementado puro	546,60	482617	1,27	2,30	0,64
Modelo Rubin	507,97	442800	1,30	2,54	0,64
Modelo erro máximo	372,96	420886	1,72	3,04	0,57

Na tabela 13, é possível notar que o modelo complementado puro e o modelo de Rubin reproduzem as estatísticas da variável U (ppm) do conjunto Walker lake original (que possui U (ppm) sub amostrado) de maneira mais satisfatória que o modelo erro máximo. Os erros das estatísticas das distribuições para o modelo Rubin são menores que 10%, assim como para o modelo complementado puro com exceção da média da distribuição que possui um desvio relativo de 21%. O modelo erro máximo possui desvios relativos das estatísticas acima de 20%, com exceção da variância que é de 10%. Contudo, a correlação entre a variável U e V é mais bem reproduzida pelo modelo erro máximo.

Os cenários médios dos modelos complementados serão utilizados como entrada para a simulação sequencial gaussiana, a fim de obter um modelo de toda a área para a variável U (ppm). O procedimento de simulação neste estudo segue os passos padrão da metodologia geoestatística, ou seja, será aplicada a transformação *nscore* (Bliss, 1934; Verly, 1984; Deustch e Journel, 1998) aos dados de entrada e a continuidade espacial será modelada.

Ao fim desta etapa haverão 3 modelos simulados, o do cenário complementado puro, Rubin e erro máximo. Na etapa subsequente, será realizada uma cossimulação sequencial gaussiana no conjunto de dados Walker lake original, onde a variável U (ppm) é sub amostrada, após os quatro modelos desenvolvidos serão comparados ao conjunto Walker lake exaustivo, para fins de validação dos modelos.

4.1.2 Simulação sequencial gaussiana (SGS) dos cenários médios

Cada cenário médio foi submetido ao desgrupamento amostral por meio da metodologia de células móveis, e após normalizados utilizando a rotina do estilo Gslib (Deutsch e Journel, 1998) **nscoremv.exe** (Barnett, 2011). Uma vez que os dados estão normalizados, a continuidade espacial foi modelada.

Os variogramas experimentais foram calculados utilizando 10 espaçamentos com tamanho de 20 metros e tolerância de 10 metros. A tolerância angular utilizada foi de 22,5˚ e largura de banda de 10 metros. As figuras 17 a 18 mostram os variogramas experimentais dos três cenários médios, complementado puro, Rubin e erro máximo, com os modelos ajustados.

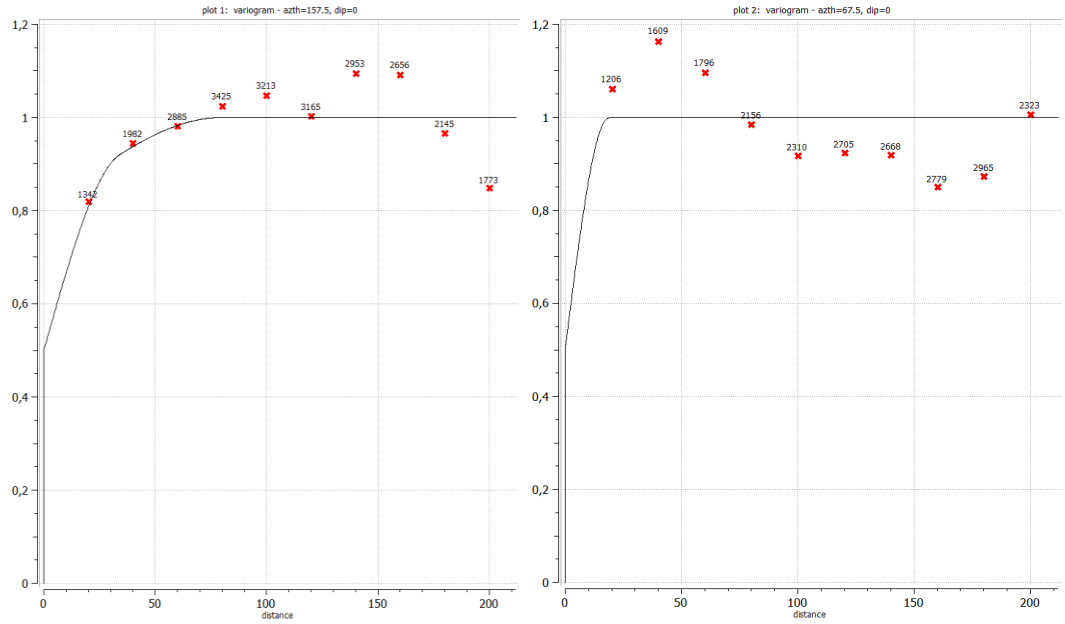

Figura 15 - Variogramas experimentais do cenário médio do modelo complementado puro. Direção de maior continuidade N157 (esquerda) e menor continuidade N67 (direita)

Na figura 15, os variogramas experimentais do cenário médio do modelo complementado puro tiveram o modelo ajustado nas direções N157, de maior continuidade e na N67, menor continuidade espacial. O modelo possui efeito pepita no valor de 0,5 além de duas estruturas esféricas, a primeira com contribuição de 0,3 da variância e a segunda com contribuição de 0,2. Os alcances são de 34m e 80m respectivamente na direção N157. Já na direção N67 os alcances são de 18m e 20m respectivamente.

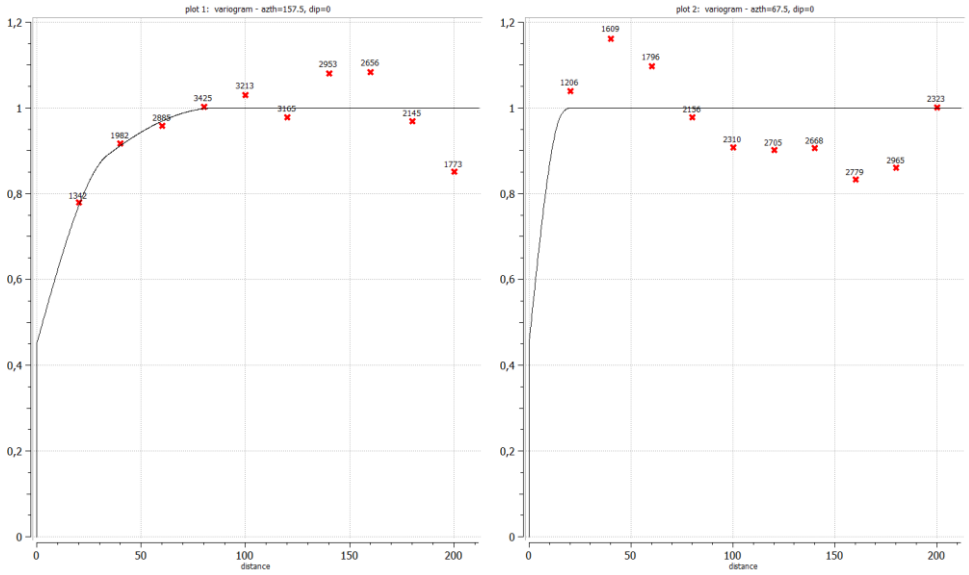

Figura 16- Variograma experimental do cenário médio do modelo Rubin, na direção de maior continuidade, N157 (esquerda), e menor continuidade, N67 (direita)

O modelo ajustado ao variograma experimental apresentado na figura 16 possui efeito pepita no valor de 0,45 além de duas estruturas esféricas, a primeira com contribuição de 0,3 e a segunda com contribuição de 0,25. Os alcances de cada estrutura são 34m e 86m respectivamente na direção de maior continuidade, N157, e 16m e 20m respectivamente na direção de menor continuidade, N67.

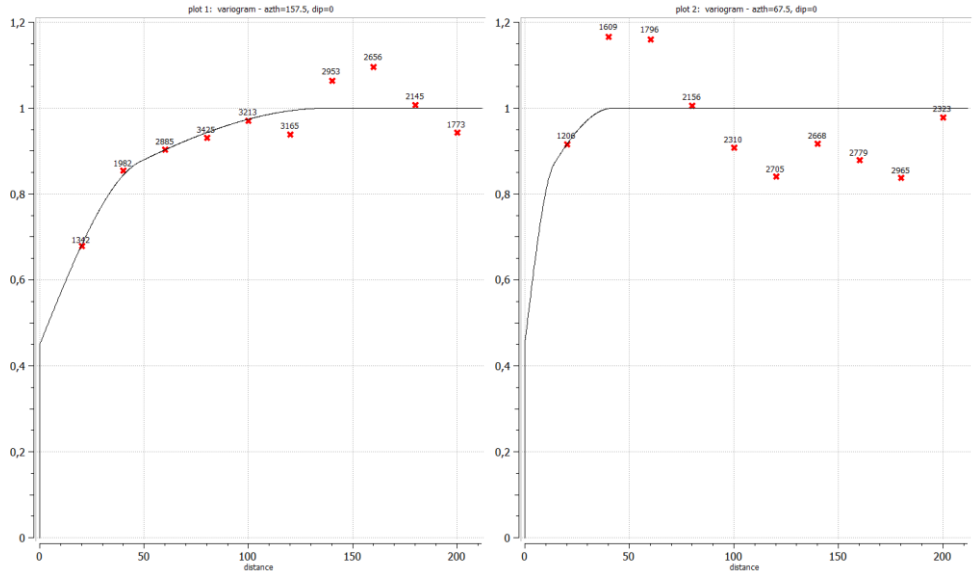

Figura 17 - Variograma experimental do cenário médio do modelo erro máximo, na direção de maior continuidade, N157 (esquerda), e na direção de menor continuidade, N67 (direita)

O variogramas experimentais do cenário médio do modelo erro máximo (figura 17) assim como os anteriores foram ajustados nas direções N157, de maior continuidade, e N67 de menor continuidade. Este modelo possui o mesmo efeito pepita e contribuições que o modelo Rubin, porém seus alcances são 48m na primeira estrutura esférica e 138m na segunda estrutura esférica na N157. Na direção N67, os alcances são 14m para a primeira estrutura esférica e 42m para a segunda estrutura esférica do modelo.

A simulação sequencial gaussiana foi aplicada a cada cenário médio utilizando os respectivos modelos de continuidade espacial, além disso a estratégia de busca de amostras para cada simulação realizada foi executada de acordo com os modelos variográficos ajustado. A tabela 14 apresenta os parâmetros utilizados na simulação sequencial gaussiana.

Tabela 14 - Parâmetros da simulação sequencial gaussiana para cada modelo gerado, complementado puro, Rubin e erro máximo.

	Modelo puro	Complementado	Modelo Rubin	Modelo erro máximo
Número de realizações	50		50	50
Mínimo de amostras	8		8	8
Máximo de dados vindos de nós simulados	12		12	12
Raio de busca	80m / 20m		86m / 20m	138m / 42m
Orientação do elipsóide de busca	157.5 / 67.5		157.5 / 67.5	157.5 /67.5
Tipo de krigagem	Simples		Simples	Simples

4.1.3 – Cossimulação sequencial gaussiana (COSGS) conjunto Walker lake

Quando o atributo de interesse é sub amostrado em relação à outras variáveis do conjunto, porém possui correlação com as demais, a abordagem recomendada é cossimulação sequencial gaussiana (COSGS). Nesse caso, não há a necessidade de isotopia amostral e, portanto, dispensa a completude dos dados por meio de inserção como a apresentada nas seções anteriores desta tese.

Apesar disto, a abordagem de COSGS apresenta algumas desvantagens: a primeira consiste na modelagem simultânea das covariâncias diretas e cruzadas por meio do modelo linear de corregionalização. Este procedimento é laborioso e por vezes, impraticável quando a

análise envolve mais de três vaiáveis. Além disto, caso o mecanismo de falta não seja MAR o modelo pode levar a resultados com viés, uma vez que a amostra faltante depende do seu próprio valor.

Nesta seção, o conjunto de dados Walker lake, possui duas variáveis U (ppm) e V (ppm), onde a primeira possui 195 amostras a menos que a segunda, será cossimulada. O modelo resultante será comparado tanto com os modelos simulados da seção anterior bem como o conjunto de dados Walker Lake exaustivo a fim de avaliar a qualidade do modelo construído.

O objetivo da cossimulação é obter um modelo para o atributo de interesse, condicionando tanto as informações conhecidas deste quanto da variável secundária, V (ppm).

Assim como a SGS, a COSGS inicia com a transformação do conjunto de dados em um novo conjunto normalizado e então a construção do modelo de continuidade espacial conjunto é feito por meio dos variogramas experimentais diretos e cruzados.

Para o conjunto em questão, os modelos foram ajustados com base na variável V (ppm), que possui mais amostras e continuidade melhor definida. Os variogramas experimentais são apresentados nas figuras 18 a 20.

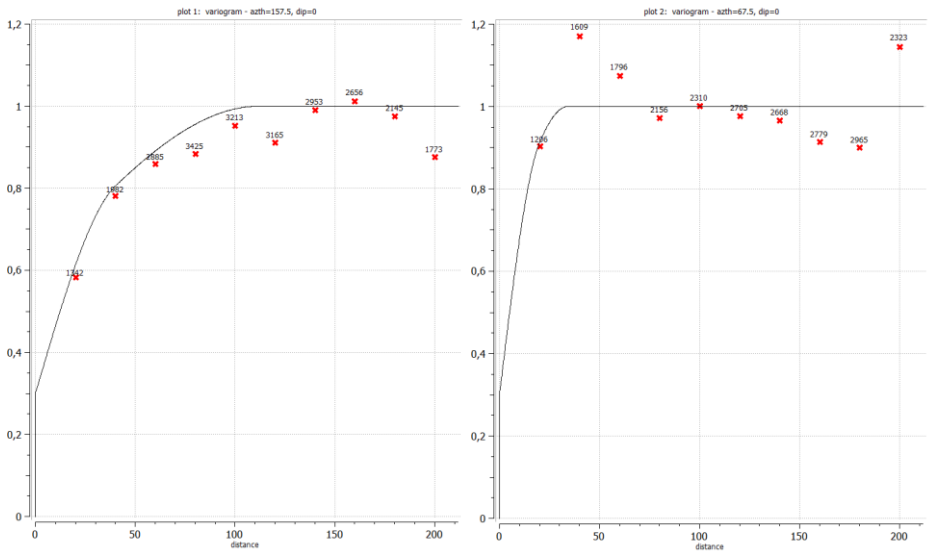

Figura 18 – Variograma experimental direto da variável V (ppm). Direção de maior continuidade: N157 (esquerda); direção de menor continuidade: N67 (direita)

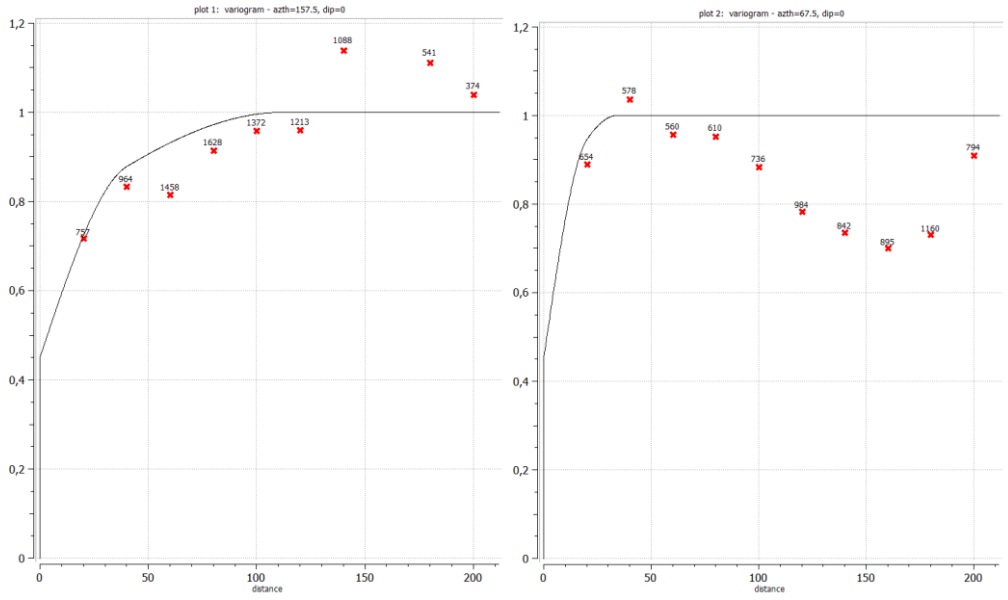

Figura 19 - Variograma experimental direto da variável U (ppm). Direção de maior continuidade: N157 (esquerda); direção de menor continuidade: N67 (direita)

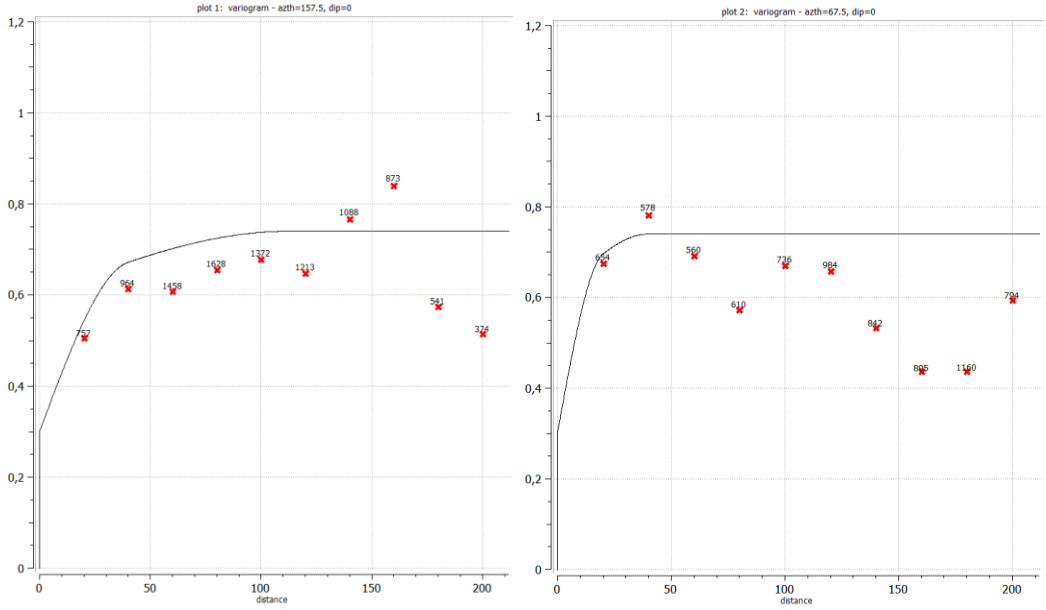

Figura 20 - Variograma experimental cruzado das variáveis U (ppm) e V (ppm). Direção de maior continuidade: N157 (esquerda); direção de menor continuidade: N67 (direita)

O modelo ajustado ao variogramas experimental possui para cada caso três estruturas. A primeira sendo o efeito pepita e as outras duas esféricas. Os alcances na direção de maior continuidade espacial, N157, são de 40m na primeira estrutura esférica e 114m na segunda estrutura esférica. Já na direção de menor continuidade, N67, os alcances são de 20m na primeira estrutura esférica e 34m na segunda estrutura esférica. As contribuições de cada estrutura do modelo são apresentadas em forma de matrizes a seguir.

$$C_0 = \begin{bmatrix} 0{,}45 & 0{,}3 \\ 0{,}3 & 0{,}3 \end{bmatrix}$$

Onde o termo $C_{0_{11}}$ corresponde ao termo do modelo direto de U(ppm), o termo $C_{0_{22}}$ ao modelo direto de V (ppm) e os termos da diagonal secundária ao modelo cruzado U (ppm) e V(ppm). A matriz C_0 corresponde a matriz do efeito pepita. As matrizes C_1 e C_2 correspondem às matrizes de contribuições das primeira e segunda estruturas esféricas respectivamente.

$$C_1 = \begin{bmatrix} 0{,}3 & 0{,}3 \\ 0{,}3 & 0{,}3 \end{bmatrix}$$

$$C_2 = \begin{bmatrix} 0{,}25 & 0{,}14 \\ 0{,}14 & 0{,}4 \end{bmatrix}$$

A partir das matrizes acima é possível ver que o modelo linear de corregionalização é satisfeito e, portanto, a cossimulação pode ser aplicada aos dados. Os parâmetros da cossimulação são apresentando na tabela 15.

Tabela 15 – Parâmetros de cossimulação sequencial gaussiana do conjunto Walker Lake.

	Cossimulação sequencial gaussiana
Número de realizações	50
Mínimo de amostras	1
Busca por setores/ número de setores	Sim/4
Máximo de amostras por setor	3
Máximo de dados vindos de nós simulados	12
Raio de busca	114m / 34m
Orientação do elipsoide de busca	157.5 / 67.5
Tipo de krigagem	Simples

4.2 Análise dos resultados

O conhecimento do conjunto de dados exaustivo, com 78000 amostras de V (ppm) e U (ppm) permite que avaliação da qualidade dos modelos gerados por meio de cada metodologia aplicada. Como parte da validação dos resultados é sabido que os modelos simulados devem reproduzir os histogramas dos dados em análise. A figura 23 apresenta um comparativo dos histogramas dos modelos simulados em relação ao histograma exaustivo

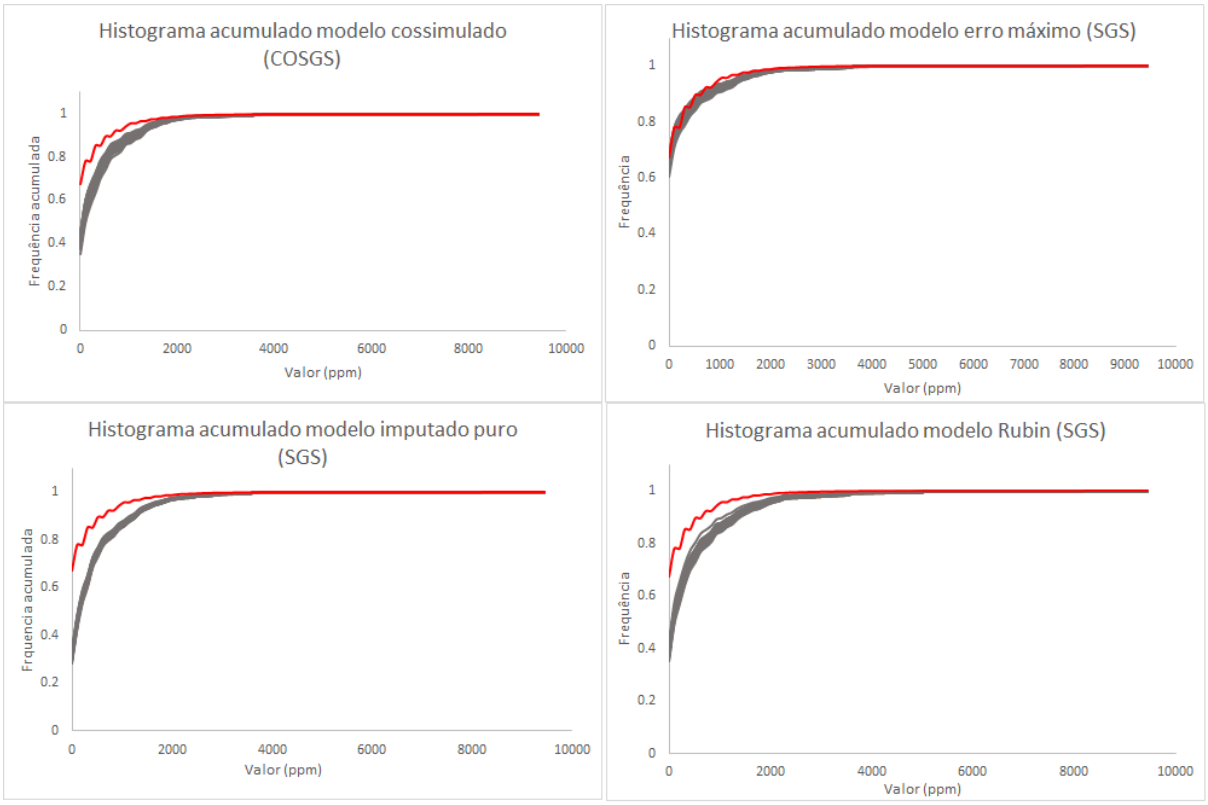

Figura 21 – Histogramas acumulados dos modelos simulados, curvas cinzas, e histograma acumulado do conjunto Walker Lake exaustivo, curva vermelha

É possível verificar, a partir da figura 21 que o modelo cossimulado e os modelos complementado puro e Rubin subestimam a frequência de valores na cauda inferior da distribuição enquanto o modelo erro máximo reproduz de maneira mais satisfatória a distribuição do conjunto exaustivo dos dados.

A tabela 16 apresenta o intervalo da estatística dos modelos simulados, bem como os modelos médios em comparação aos dados exaustivos.

Tabela 16 - comparativo entre os intervalos de valores obtidos para algumas estatísticas nos 50 modelos simulados a partir de cada metodologia bem como os cenários médios de cada modelo simulado e conjunto de dado referência, Walker Lake exaustivo.

Modelo	Média	Variância	Assimetria
Dados exaustivos	266,0	238585	3,72
COSGS (50 realizações)	347,9 – 453,5	294852 - 421190	2,74 – 3,20
SGS erro máximo (50 realizações)	226,2 – 315,0	269536 - 396347	3,53 – 4,2
SGS complementado puro (50 realizações)	454,2 – 515,7	365277 – 460684	2,42 – 2,71
SGS Rubin (50 realizações)	398,1 – 535,2	407922 – 637572	2,69 – 3,21
COSGS médio	402,99	361136,36	2,79
SGS erro máximo médio	268,63	333378,22	3,78
SGS complementado puro médio	484,76	399868,2	2,22
SGS Rubin médio	475,76	534209,16	2,91

Note na tabela 16 que os intervalos dos valores para as estatísticas das distribuições nos modelos simulados apresentam viés sistemático em relação à estatística do conjunto exaustivo.

Para simulados por meio de COSGS, SGS Rubin e SGS complementado puro a média global é superestimada em todas as realizações, bem como a variância. A assimetria por sua vez, nos modelos por meio de COSGS, SGS Rubin e SGS complementado puro é subestimada em todas as realizações, constituindo neste caso também um viés sistemático em relação a este parâmetro.

No modelo SGS erro máximo, a média global do conjunto exaustivo está contida no intervalo de valores obtidos nas 50 realizações, apresentando um desvio relativo no modelo simulado médio de 0,97% em relação ao valor do conjunto exaustivo. Além disso, o valor de assimetria do conjunto exaustivo neste caso também está contido no intervalo de valores obtidos nas 50 realizações do modelo erro máximo, de modo que o desvio do valor obtido no cenário médio em relação ao valor real é 1,61%. Todavia, a variância neste caso também foi superestimada em todos os cenários realizados.

É possível perceber que o modelo simulado por meio de SGS, onde foi aplicada a transformação fixa do erro máximo, apresentou melhoras no modelo resultante quando confrontado com os demais modelos gerados.

A seguinte validação é a reprodução da média local por meio de análise de deriva do conjunto simulado utilizando o conjunto Walker lake exaustivo como referência.

A área de estudo foi dividida em faixas de 13m na direção leste e 15m na direção norte, a média do conjunto exaustivo e dos modelos simulados é calculada dentro destas faixas e o comportamento simultâneo da referência e do modelo é analisando. A figura 22 apresenta o gráfico que sumariza tal comportamento em ambas direções para os modelos simulados.

Figura 22 - Análise de deriva dos modelos simulados por meio de Simulação sequencial gaussiana (SGS) e cossimulação sequencial gaussiana (COSGS), curvas cinzas, em relação ao conjunto referência Walker Lake exaustivo (curva vermelha)

Nota-se na figura 22 que o modelo complementado puro simulado por meio de SGS superestima os valores da média tanto na direção leste quanto na direção norte em todos os locais, exceto na faixa de 50m na direção leste, mostrando neste gráfico que o modelo não obteve valores coerentes à realidade do depósito. A amostragem preferencial de valores altos da distribuição e o desprezo da distribuição na cauda inferior impactam o modelo gerado, que apresenta viés claro em suas inferências. Assim, também é o comportamento do modelo Rubin, de modo que a transformação fixa não se mostra suficiente para corrigir o viés nas amostras transmitidos para o modelo.

O modelo COSGS por sua vez, também apresenta viés no modelo simulado resultante, onde é possível observar da figura 22 que há regiões em que todas as realizações do modelo

apresentam valores mais altos que os do conjunto referência (linha vermelha). Porém, neste estudo de caso, o viés gerado por meio da cossimulação das variáveis, utilizando o modelo linear de corregionalização, é menos acentuado que os modelos simulados utilizando a inserção pura e a transformação fixa de Rubin. Ainda, é possível que o modelo cossimulado obtivesse um viés menos significativo caso o desagrupamento amostral fosse realizado por meio de polígonos de Voronoi (Dirichlet, 1850) retendo os pesos designados, ainda que não haja garantia da que eliminação completa do viés resultante da amostragem preferencial de U (ppm), uma vez que os valores utilizados neste procedimento continuariam sendo todos valores U (ppm) pertencentes a regiões superiores da distribuição do conjunto exaustivo. Contudo, caso a cossimulação fosse condicionada a uma distribuição representativa do conjunto exaustivo essa também apresentaria resultados sem viés presente.

Por fim, é possível ver na figura 22 que a transformação fixa utilizando o erro máximo foi capaz de corrigir o viés originalmente presente no modelo complementado puro, de maneira que os valores do conjunto referência estão contidos dentro da nuvem de realizações do modelo simulado, não apresentando viés no modelo final gerado.

Foi avaliado também o espaço de incerteza das simulações por meio de um gráfico de acuracidade, uma vez que se tem o conjunto de dados exaustivo como referência. A figura 23 apresenta o gráfico de acuracidade para todos os modelos simulados.

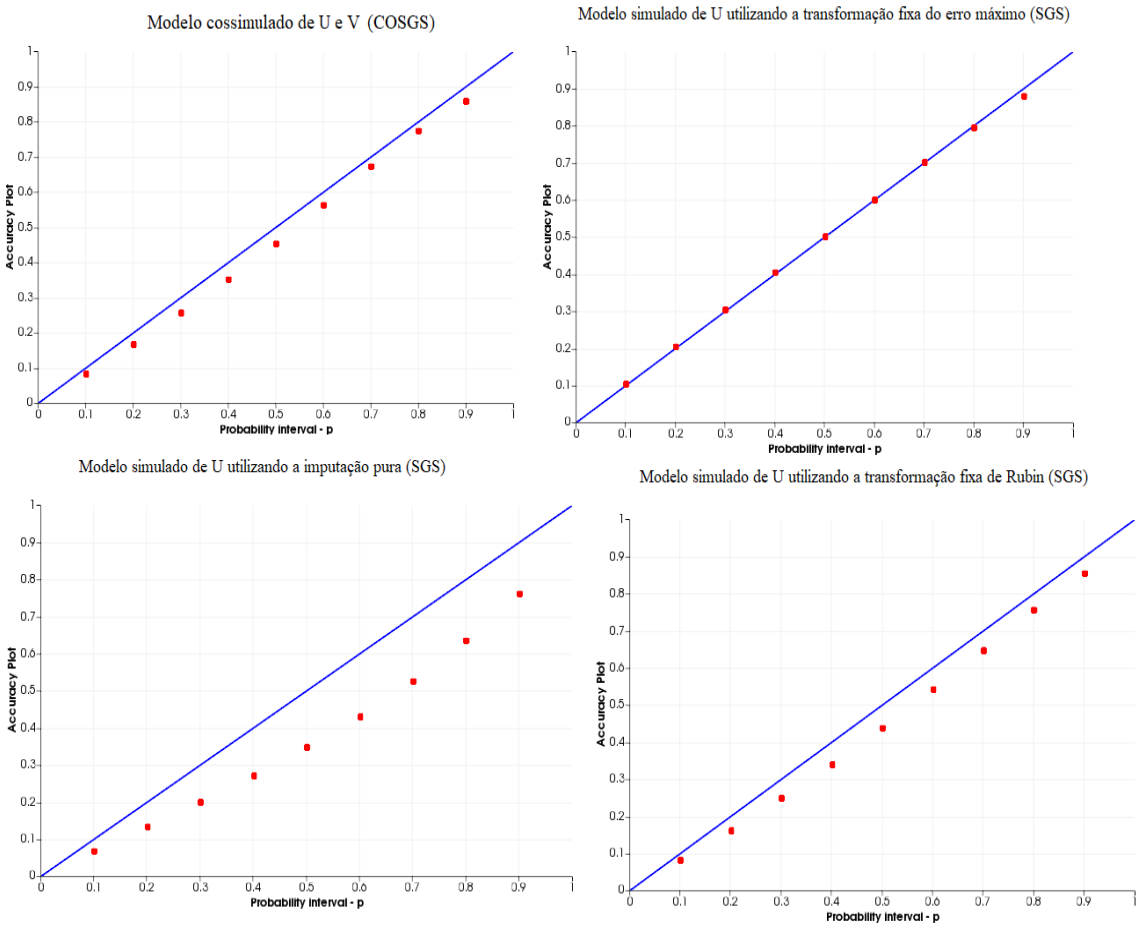

Figura 23 - Gráficos de acuracidade para os modelos simulados da variável U (ppm)

Em gráficos de acuracidade, pontos que se localizam acima da reta de 45˚ são considerados acurados, ponto que se localizam abaixo da reta de 45˚ são considerados inacurados, e pontos que estão na reta de 45˚ são precisos e acurados. Vemos que os modelos COSGS, complementado puro e Rubin são ligeiramente inacurados, ou seja, há discrepâncias entre os valores de probabilidade inferidos pelos modelos e os valores reais no conjunto exaustivo. Contudo, os modelos COSGS e Rubin apresentam melhor desempenho que o modelo simulado do cenário complementado puro, onde não foi realizada nenhuma transformação fixa aos valores complementados e por consequência estes resultam em um modelo que possui viés mais evidente em relação ao conjunto real de dados. O modelo simulado em que a transformação fixa do erro máximo foi aplicada, por sua vez, possui precisão em todos os intervalos de probabilidade, exceto o intervalo de probabilidade 90% onde pode-se perceber ligeira inacurácia do modelo proposto.

Os resultados obtidos mostram que neste estudo de caso, o tratamento de dados faltantes por meio de inserção e aplicação da transformação fixa do erro máximo aos dados

complementados resultou em evidentes melhoras do modelo simulado gerados a partir do conjunto de dados.

Vale ressaltar, que como foi recomendado no capitulo 4 desta tese e baseado nos resultados obtidos no estudo de caso apresentado neste capitulo, conjuntos de dados que apresentem altos coeficientes de variação e assimetria da distribuição significativa são quadros adequados a transformação fixa do erro máximo. Mostra-se nestes casos, que os modelos resultantes obtiveram estatística e precisão mais satisfatórias que as outras abordagens adotadas neste estudo de caso.

Além disso, é importante ressaltar que mesmo o conjunto sendo multivariado, a simulação sequencial gaussiana do cenário complementado e transformado pelo erro máximo (univariadas) obteve melhores resultados que o modelo cossimulado utilizando o modelo linear de corregionalização. Este fato permite que mesmo que não sejam aplicadas metodologias de fatoração de variáveis correlacionadas, a correção da distribuição condicionante da simulação previamente ao processo permite melhora o modelo final elaborado.

Capitulo 5

O problema proposto

No capítulo 1 desta tese, foi exposto o problema que motivou o desenvolvimento do estudo explorado neste documento. Foram colocados os problemas existentes no que se refere a conjuntos de dados multivariados, bem como a influência que a ausência de amostras em algumas variáveis do conjunto pode exercer sobre o modelo resultante gerado. Foi esclarecido no capítulo 1, que as abordagens geoestatísticas que são capazes de lidar com conjuntos multivariados heterotópicos impõem o uso do modelo linear de corregionalização, uma vez que a heterotopia impede a aplicação de metodologias de fatoração como PCA, MAF e PPMT.

Além disso, foi exposto que se o mecanismo de falta atuante no conjunto fosse MNAR, a aplicação de metodologias de coestimativa, ainda que utilizando a modelagem simultânea das covariâncias diretas e cruzadas por meio do modelo linear de corregionalização, levaria às saídas enviesadas e não representativas do fenômeno em análise, uma vez que a distribuição do conjunto de dados que alimenta o modelo está distorcida devido à falta amostral.

5.1 Conclusões Iniciais

A meta proposta inicialmente no item 1.5 do capitulo 1, de obter uma metodologia de fácil implementação computacional que mitigue o viés gerado ao modelo devido ao mecanismo MNAR e considere o comportamento espacial do conjunto de dados foi parcialmente atingida pela abordagem proposta nesta tese. Nos estudos de caso apresentados, os resultados mostram-se satisfatórios. Porém, a reprodução estatística não é completa, uma vez que os modelos gerados por meio de inserção falham ao reproduzir a variabilidade dos dados exaustivos. Contudo, a metodologia clássica, COSGS também falha neste aspecto e desta forma a abordagem proposta se mostra tão eficiente quanto a abordagem clássica com menor custo laboral e computacional.

5.2 Conclusões referentes aos objetivos específicos

No que se refere aos primeiros objetivos propostos nesta tese, a fim de investigar e adaptar metodologias de inserção que fossem adequadas ao problema de dados faltantes sob mecanismo MNAR foi atingido, uma vez que diversas metodologias passíveis de aplicação ao

problema foram expostas ao longo do capitulo 2 e, ao longo do capitulo 3 foi explicitada a metodologia que combinou as abordagens de atualização bayesiana (Doyen, 1996; Ren, 2007) e transformações fixas (Rubin,1987), propondo uma nova métrica a ser utilizada. Métrica essa, que no estudo, é calibrada pelo próprio erro de inserção assumindo o mecanismo MAR.

Em relação ao terceiro objetivo proposto, realizar a análise comparativa entre diferentes métricas de aplicação da transformação fixa, este também foi atingido. Ao longo dos capítulos 3 e 4 deste documento, são expostos os modelos gerados sem aplicação de transformação fixa aos valores complementados, bem como os modelos em que foi aplicada a transformação fixa proposta por Rubin (1987), de 20% do valor complementado e a transformação fixa proporcional ao erro máximo obtido no modelo de inserção por atualização bayesiana. Foi visto, em ambos estudos de caso apresentados, capítulos 3 e 4, que os modelos são beneficiados pela aplicação das transformações fixas aos valores complementados, onde distribuições dos modelos tornaram-se mais representativas do conjunto de dados completos do que a obtida por meio do modelo MAR de atualização bayesiana sem qualquer correção *a posteriori* das saídas dos modelos.

Foi verificado, nos estudos de caso aqui apresentados, o benefício da utilização de modelos facilmente comunicáveis, de modo que a partir do modelo gerado por meio da atualização bayesiana, rapidamente modelos com saídas MNAR podem ser construídos e avaliados. Assim, foi visto nos estudos apresentados, que para variáveis fortemente assimétricas, a metodologia que transforma os valores complementados por meio do erro máximo obteve melhores resultados que os outros modelos gerados por meio das diversas metodologias utilizadas.

Em relação a construção de modelos estocásticos por meio de SGS dos conjuntos de dados complementados, conclui-se que a utilização do cenário médio dos conjuntos complementados não adiciona viés ao modelo resultante.

Na comparação dos modelos obtidos por meio de SGS dos cenários médio e COSGS do conjunto de dados heterotópico, foi visto que o modelo gerado por meio de SGS dos dados transformados pela métrica do erro máximo, obteve resultados mais coerentes com a realidade do conjunto exaustivo utilizado como referência. Já o modelo em que foi aplicada a transformação fixa de Rubin apresentou resultados equivalentes ao obtido por meio de COSGS. O modelo estocástico construído a partir do cenário médio complementado por atualização

bayesiana sem qualquer transformação fixa aplicada obteve performance inferior ao modelo COSGS.

Foi verificado que a abordagem clássica, COSGS, apresentou viés em seu modelo final, mostrando não ser capaz de corrigir a distorção da distribuição gerada pela falta MNAR de dados. Além disso, a inserção por meio de atualização bayesiana, sendo uma metodologia desenvolvida para falta MAR, foi verificado que esta não é capaz de mitigar o viés gerado pela falta de dados MNAR.

Assim, considerando os estudos de caso realizados nesta tese, a aplicação das transformações fixas, em especial a métrica do erro máximo de inserção, gera ganhos nos modelos resultantes. Esta produziu um modelo mais representativo dos dados e dispensou a modelagem simultânea das covariâncias diretas e cruzadas das variáveis envolvidas. Ainda, os modelos gerados são facilmente intercambiáveis, permitindo que o usuário teste inúmeras transformações fixas ao conjunto de dados complementados e avalie a conformidade dos resultados obtidos.

É importante ressaltar, que para aplicação da metodologia proposta, é necessário que o analista tenha conhecimento a respeito do comportamento dos atributos geológicos em estudo e da estratégia amostral, de maneira que este possa avaliar a melhor abordagem a ser utilizada em seu conjunto de dados.

5.3 Recomendações para trabalhos futuros

A metodologia apresentada nesta tese não é única e abre espaço para outras formas de acessar o problema de faltantes MNAR. Como trabalhos futuros é de interesse que outras métricas para transformações fixas sejam exploradas, e que a aplicação da metodologia tenha sua validade checada em conjunto de dados de diferentes origens por exemplo, dados geometalúrgicos.

Outra linha de estudo recai sobre o impacto que dados faltantes MNAR tem em conjuntos de dados composicionais e a influência a falta amostral não aleatória exerce sobre o fechamento estequiométrico no modelo final gerado, além da aplicação da metodologia explorada nesta tese a conjunto de dados desta natureza.

Também se sugere, que a aplicação seja realizada em um conjunto de dados em que mais de duas variáveis sejam faltantes, além da aplicação da metodologia combinada a métodos de fatoração como PCA, MAF e PPMT.

Recomenda-se que sejam exploradas maneiras de utilizar somente as distribuições geradas pela inserção combinada a transformações fixas (por exemplo condicionar simulações geoestatísticas) em oposição ao uso das amostras imputadas como *hard data*.

Além disso, se indica que os modelos de mistura de padrões e de seleção devem ser explorados e, que uma análise de sensitividade deve ser realizada nos modelos resultantes das duas metodologias citadas, uma vez que tais abordagens não foram testadas em dados geológicos.

REFERÊNCIAS BIBLIOGRÁFICAS

Almeida, A. S., 1993, JOINT SIMULATION OF MULTIPLE VARIABLES WITH A MARKOV-TYPE COREGIONALIZATION MODEL, Tese de Doutorado, Stanford University, 199p.

Barnett, R. M. e Deutsch, C. V., 2012, IMPUTATION OF GEOLOGIC DATA, Reporte anual, University of Alberta, 16p.

Barnett, R. M., Manchuck, J. G. e Deutsch, C. V., 2014, PROJECTION PERSUIT MULTIVARIATE TRANSFORM, Mathematical Geoscience, 46 (3), pp. 337-359.

Bliss, C. L., 1934, THE METHOD OF PROBITS, Science, 79 (2037), pp. 38-39.

Box, G. E. P. e Tiao, G. C., 1973, BAYESIAN INFERENCE IN STATISTICAL ANALYSIS, John Wiley and Sons Ltda, 608p.

Burkett, J. P., 2009, BAYESIAN ANALYSIS AND MULTIPLE PARAMETER MODELS, Notas de Aula, University of Rhode Island, 22p.

Cohen, J., 1988, SATATISTICAL POWER ANALYSIS FOR BEHAVIOURAL SCIENCES, Erlbaum publishers, 559p.

Deutsch C. V. e Journel, A. G., 1998, GSLIB: A GEOSTATISTICAL SOFTWARE LIBRARY AND USER'S GUIDE, Oxford University Press, 363p.

Dempster, A. P., Laird, N. M. e Rubin, D. B., 1977, MAXIMUM LIKELIHOOD FROM INCOMPLETE DATA VIA THE EM ALGORITHM, Journal of the Royal Statistical Society, 39 (1), pp. 1-38.

Davis, B. e Greenes, K. A., 1983, ESTIMATION USING SPATIALLY DISTRIBUTED MULTIVARIATE DATA: AN EXAMPLE WITH COAL QUALITY, Mathematical Geology, 15 (2), pp. 287-300.

Desbarats, A., 2001, GEOSTATISTICAL MODELING OF REGIONALIZED GRAIN SIZE DISTRIBUTIONS USING MIN/MAX AUTOCORRELATION FACTORS, Proceedings of the third European conference on geostatistics for environmental applications, pp. 441-452.

Desbarats, A. e Dimitrakopoulos, R., 2000, GEOSTATISTICAL SIMULATION OF REGIONALIZED PORE-SIZE DISTRIBUTIONS USING MIN/MAX AUTOCORRELATION FACTORS, Mathematical Geology, 32 (8), pp. 919-942.

Dixon, W. J., 1988, BMDP STATISTICAL SOFTWARE, University California Press, 619p.

Doyen P. M., Den Boer L. D. e Pillet, W. R., 1996, SEISMIC POROSITY MAPPING IN THE EKOFISK FIELD USING A NEW FORM OF COLLOCATED COKRIGING, SPE Annual Technical Conference and Exhibition, pp. 11-21.

Enders, C. K., 2010, APPLIED MISSING DATA ANALYSIS, The Guilford Press, 401p.

Dirichlet, G. L., 1850, ÜBER DIE REDUKTION DER POSITIVE QUADRATISKEN FORMEM MI DREI UNBESTIMMTEN GANZEN ZAHLEN, Journal für die reine und angewandte Mathematik, pp. 209-227.

Goovaerts, P., 1997, GEOSTATISTICS FOR NATURAL RESOURCE EVALUATION, Oxford University Press, 483p.

Hartman, R. S., 1991, A MONTE CARLO ANALYSIS OF ALTERNATIVE ESTIMATORS IN MODELS INVOLVING SELECTIVITY, Journal of Business and Economic Statistics, 9 (1), 41–49.

Heckmann, J. T., 1979, SAMPLE SELECTION BIAS AS A SPECIFICATION ERROR, Econometrica, 47 (1), pp. 153-161.

Hoettling, H., 1933, ANALYSIS OF A COMPLEX STATISTICAL VARIABLES OF PRINCIPAL COMPONENTS, Journel of educational psycolagy, 24 (7), pp. 417-441.

Isaaks, E. H., 1990, THE APPLICATION OF MONTE CARLO METHODS TO THE ANALYSIS OF SPATIALLY CORRELATED DATA, Tese de doutorado, Stanford University, 213p.

Isaaks, E. H. e Srivastava, R. M., 1989, AN INTRODUCTION TO APPLIED GEOSTATISTICS, Oxford University Press, 561p.

Jollife, I. T., 2002, PRINCIPAL COMPONENT ANALYSIS, Springer, 518p.

Journel, A.,1999, MARKOV MODEL FOR CROSS-COVARIANCES, Mathmatical Geology, 31 (8), pp. 955-964.

Kim, K. H. e Bentler, P. M., 2002, TESTS OF HOMOGEINITY OF MEANS AND COVARIANCE MATRICES FOR MULTIVARIATE INCOMPLETE DATA, Psychometrika, 67 (4), pp. 609-624.

Lauanthong, O. e Deutsch, C. V., 2003, STEPWISE CONDITIONAL TRANFORMATION FOR SIMULATION OF MULTIPLE VARIABLES, Mathematical Geology, 35 (2), pp. 155-173.

Little, R. J. A., 1988, A TEST OF MISSING COMPLETELY AT RANDOM FOR MULTIVARIATE DATA WITH MISSING VALUES, Journal of the American Statistical Association, 83 (404), pp. 198-1202.

Little, R. J. A., 1993, PATTERN-MIXTURE MODELS FOR MULTIVARIATE INCOMPLETE DATA, Journal of the American Statistical Association, 88 (421), pp. 125-134.

Little, R. J. A. e Rubin, D. B., 2002, STATISTICAL ANALYSIS WITH MISSING DATA, John Wiley and Sons, 371p.

Marechal, A., 1970, COKRIEGEAGE ET REGRESSION EM CORRELATION INTRISIQUE, Centre de Geostatistique de Fontainebleau, 40p.

Matheron, G., 1962, TRAITÉ DE GÉOSTATISTIQUE APPLIQUÉE, Technip, 333p.

Matheron, G.,1963, PRINCIPLES OF GEOSTATISTICS, Economic Geology, 58 (8), pp. 1246-1266.

Nawata, K., 1993, A NOTE ON THE ESTIMATION OF MODELS WITH SAMPLE-SELECTION BIASES, Economics Letters, 42 (1),15–24.

Pearson, K., 1901, ON LINES AND PLANES CLOSEST FIT TO SYSTEMS OF POINTS IN SPACE, Phylosophical Magazine, 2 (6), pp. 559-572.

Puhani, P. A., 2000, THE HECKMAN CORRECTION FOR SAMPLE SELECTION AND ITS CRITIQUE, Journal of Economic Surveys, 14 (1), 53–67.

Ren, W., 2007, EXACT DOWNSCALING IN RESERVOIR MODELING, Tese de Dourtorado, University of Alberta, 213p.

Reuwsaat, J. D. V., 2011, ESTUDO DE ALTERNATIVAS DE ESTIMATIVA PARA JAZIDAS DE FERRO UTILIZANDO DADOS HETEROTOPICOS, Dissertação de Mestrado, Universidade Federal do Rio Grande do Sul, 180p.

Rondon, O., 2011, TEACHING AID: MINIMUM/MAXIMUM AUTOCORRELATION FACTORS FOR JOINT SIMULATIONS OF ATTRIBUTES, Mathematical Geosciences, 44 (4), pp. 469-504.

Rosemblat, M., 1952, REMARKS ON A MULTIVARIATE TRANSFORMATION, Annals of Mathematical statistics, 23 (3), pp. 470-472.

Rubin, D. B., 1976, INFERENCE AND MISSING DATA, Biometrika, 63 (3), pp. 51-592.

Rubin, D. B., 1987, MULTIPLE IMPUTATION FOR NONRESPONSE IN SURVEYS, John Wiley and Sons Ltda, 253p.

Schafer, J. L., 2003, MULTIPLE IMPUTATION IN MULTIVARIATE PROBLEMS WHEN THE IMPUTATION AND ANALYSIS MODELS DIFFER, Statistica Neerlandica, 57 (1), pp. 19-35.

Schafer. J. L. e Graham, J. W., 2002, MISSING DATA: OUR VIEW OF THE STATE OF THE ART, Psychological Methods, 7 (2), pp. 147-177.

Silva, C. Z., 2013, A DESCORRELAÇÃO DE VARIÁVEIS COM FATORIZAÇÃO MAF EM ESTIMATIVA DE TEORES, Dissertação de Mestrado, Universidade Federal do Rio Grande do Sul, 142p.

Switzer, P. e Green, A. A., 1984, MIN/MAX AUTOCORRELATION FACTORS FOR MULTIVARIATE SPATIAL IMAGERY, Reporte Técnico Número 6, Universidade de Stanford, 14p.

Verly, G. W., 1993, SEQUENTIAL GAUSSIAN COSSIMULATION: A SIMULATION METHOD INTEGRATING SEVERAL TYPES OF INFORMATION, In: Sores, A., Geostatistics Tróia ’92. Qantitative Geology and Geostatistics, Springer, pp. 543-554.

Verly, G. W., 1984, ESTIMATION OF SPATIAL POINT AND BLOCK DISTRIBUTIONS: THE MULTIGAUSSIAN MODEL, Tese de Doutorado, Stanford University, 416p.

Watanabe, J., Yamamoto, J. K., Rocha, M. M. e Fonseca, P. P., 2009, ESTUDO DA INFLUÊNCIA DA CORRELAÇÃO INICIAL ENTRE VARIÁVEIS NOS RESULTADOS DE COESTIMATIVA, Geociencias, 28(4), pp. 467-484.

Wackernagel, H., 1998, PRINCIPAL COMPONENT ANALYSIS FOR AUTOCORRELATED DATA: A GEOSTATISTICAL PERSPECTIVE, Reporte Técnico Número 22, Ecole de Mines de Paris, 41p.

Wilks, S. S., 1932, MOMENTS AND DISTRIBUTIONS OF ESTIMATE OF POPULATION PARAMETERS FROM FRAGMENTARY SAMPLES, The annuals of Mathematical Statistics, 3 (3), pp. 163-195.

Xu, W., Tran, T. T., Srivastava, R. M. e Journel, A. G., 1992, INTEGRATING SEISMIC DATA IN RESERVOIR MODELING: THE COLLOCATED COKRIGING ALTERNATIVE, Proceedings of 67th Annual Technical Conference of the Society of Petroleum Engineers, pp. 833-842.