UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE MINAS, METALÚRGICA E DE MATERIAIS (PPGEM)

VICTOR MIGUEL SILVA

GEOESTATÍSTICA NA AUSÊNCIA DE HARD DATA:
lidando com o erro amostral

Porto Alegre
2019

VICTOR MIGUEL SILVA

GEOESTATÍSTICA NA AUSÊNCIA DE HARD DATA:

lidando com o erro amostral

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Minas, Metalúrgica e dos Materiais – PPGEM, como parte dos requisitos para a obtenção do título de Doutor em Engenharia.

Orientador: Prof. Dr. João Felipe Coimbra Leite Costa

Porto Alegre

2020

CIP - Catalogação na Publicação

Miguel Silva, Victor
GEOESTATÍSTICA NA AUSÊNCIA DE HARD DATA: lidando com o erro amostral / Victor Miguel Silva. -- 2020.
89 f.
Orientador: João Felipe Coimbra de Costa Leite.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Escola de Engenharia, Programa de Pós-Graduação em Engenharia de Minas, Metalúrgica e de Materiais, Porto Alegre, BR-RS, 2020.

1. Erro amostral. 2. Co-Krigagem. 3. Simulação Geoestatística. 4. Geoestatística. 5. Espaço de incerteza. I. Coimbra de Costa Leite, João Felipe, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

VICTOR MIGUEL SILVA

GEOESTATÍSTICA NA AUSÊNCIA DE HARD DATA:

lidando com o erro amostral

Esta tese foi julgada adequada para a obtenção do título de Doutor em Engenharia e aprovada em sua forma final pelo Orientador e pela Banca Examinadora do Curso de Pós-Graduação.

_____

Orientador: Prof. Dr. João Felipe Coimbra Leite Costa

_____

Coordenador do PPGEM: Prof. Dr. Afonso Reguly

Aprovado em: 23/04/2020

BANCA EXAMINADORA:

Camilla Zacche da Silva – University of Alberta

Luiz Eduardo Seabra Varella - Petrobras

Vanessa Cerqueira Koppe – LPM / UFRGS

À minha família,

principalmente, à Daniele.

AGRADECIMENTOS

Ao Professor João Felipe Coimbra Leite Costa, pela orientação e apoiando o meu desenvolvimento desde o início, em 2012, na CBA, onde havia um contrato de treinamento e avaliação de furos gêmeos, evoluindo para o mestrado e, agora, para o doutorado. Agradeço ao Professor Clayton Deutsch pela coorientação ao longo de 2018 e 2019.

À Daniele, pela compreensão ao longo de 6 anos de dificuldades em se conciliar às minhas carreiras profissional e acadêmica, além de todo o apoio quando decidimos nos mudar para o Canadá.

[...]The results shown here do not invalidate the Big Bang theory.

Avery (1996), em relatório interno da Organização Europeia
para Pesquisa Nuclear (CERN) sobre a definição de pesos
ótimos para minimizar o erro de estimativa de medições
com erros de medição correlacionados.

RESUMO

Na geoestatística, são chamados de *hard data* as observações do fenômeno de interesse que sejam isentas de erro ou assumidas como tal. No entanto, tal tipo de dado não pode ser obtido experimentalmente, pois o erro amostral é intrinsicamente associado a qualquer processo de amostragem. Em dados reais, erros amostrais com variância correspondendo de 10% a 40% da variância total, os dados são considerados como boas práticas ou *benchmarks, sendo,* então, comumente assumidos como isentos de erro nas rotinas geoestatísticas. A proposta do trabalho é investigar se a hipótese de que assumir em problemas geoestatísticos dados reais como *hard data* é incorreta. Estatísticas como correlação espacial, as distribuições e a estrutura de correlação medida através de observações são combinações do comportamento do fenômeno real com o dos erros e, portanto, realizações estocásticas condicionadas a honrar os parâmetros das observações não são equiprováveis ao fenômeno real. Enquanto os fluxos de trabalhos convencionais geram realizações condicionadas a honrar os parâmetros e valores dos dados, esta tese apresenta uma série de métodos que possibilitam utilizar observações afetadas por erros para gerar realizações equiprováveis ao fenômeno real. A tese é separada em cinco partes: (i) é desenvolvido um modelo de erros generalizado, tanto univariado quanto multivariado; (ii) São apresentadas alternativas para estimar o erro associado a cada medição; (iii) o covariograma e a distribuição do fenômeno real são inferidos através dos valores amostrados, dos seus erros estimados e do covariograma e distribuição ajustada ao valores amostrados. No caso multivariado, também é inferida a estrutura de correlação entre as variáveis; (iv) bancos de dados de *hard data* são gerados ao substituir as observações iniciais por simulações de possíveis valores do fenômeno real. Cada banco de dados é utilizado para simular o fenômeno de interesse em todo o domínio, sendo tanto os bancos de dados quanto as realizações do modelo condicionadas a reproduzir estatísticas inferidas do fenômeno real. Por fim, a parte (v) apresenta as conclusões e propostas de novos trabalhos. Na tese são apresentados diversos exemplos como forma de elucidar o método e demonstrar o impacto e relevância de cada etapa. Os resultados do método proposto são a geração de modelos realmente equiprováveis ao fenômeno real e espaços de incerteza que reproduzem melhor a verdadeira distância entre o modelo e a realidade.

Palavras-chave: Erro amostral; Co-Krigagem; Simulação Geoestatística; Geoestatística; Espaço de incerteza.

ABSTRACT

Sampling error with variance corresponding to 10% to 40% of the total dataset variance are considered as good practice and frequently assumed as error-free data in the geostatistical workflow. The sampling error is intrinsically associated with any sampling process. Therefore, it is impossible to obtain in practice error-free observations of the phenomenon of interest (named as hard data). This thesis investigates the hypothesis that assuming the existence of hard data in geostatistical problems is incorrect and that impacts the quality of the simulated models. Spatial correlation, distribution and the structure of correlation measured by samples combine the real behavior of the underlying-true phenomenon and the sampling error behaviour. Realizations conditioned to honor the parameters fitted to observations are not equiprobable to the underlying true phenomenon. This thesis presents a number of methods that correctly manage data affected by sampling error. The thesis is separated into five parts: (i) a generalized error model that deals either with univariate and multivariate data is developed. In the multivariate case, the error associated with observations of different variables can be correlated; (ii) alternatives are presented to estimate the error associated with each observation; (iii) the covariogram and the distribution of the underlying-true phenomenon are inferred through statistics adjusted to the observations, their estimated associated errors, and the error behaviour associated to each observation. In the multivariate case, the structure of correlation between pairs of variables is inferred; (iv) hard data cannot be sampled, but equiprobable hard data values can be simulated. Data sets are generated by replacing initial observations by simulations of hard data. All realizations generated in the simulation steps are conditioned to reproduce the inferred statistics of the underlying true phenomenon. In the last part (v), discussions ands conclusions are presented. For sake of clarity, several short examples are presented to elucidate the method and demonstrate the impact and relevance of each step. The results of the proposed method are the generation of models with a better reproduction of what is supposed to be equiprobable realizations of the underlying true process, as well as improve the simulated space of uncertainty between the model and reality.

Keywords: Sampling error; Co-kriging; Geostatistical Simulation; Geostatistics; Space of Uncertainty.

SUMÁRIO

1 INTRODUÇÃO

Amostras formam a base de informações usadas na construção de modelos geológicos na indústria mineral. São coletadas para medir uma ou mais propriedades de interesse, no entanto. As leituras dos valores associados a essas amostras, invariavelmente, possuem erros. Não existem dados livre de erro na mineração. A Teoria da Amostragem (GY, 1982) demonstra que é impossível se obter dados sem erro através de amostras extraídas de um material heterogêneo. Ainda assim, é prática comum agrupar os dados em *hard data* e *soft data*[1] em função do método de amostragem e de análise empregado ou em função de dados aprovados ou reprovados pelo QA/QC, entre outros. Mesmo amostras geradas pelas melhores práticas de amostragem e controle de qualidade levam a observações em que o coeficiente de variação do erro entre valores das duplicatas de 10 até 40% (ABZALOV, 2008).

Além do erro fundamental associado a delimitação, extração e preparação das amostras, há outras fontes de erros. Portanto, é incorreto considerar que os erros são homogêneos entre amostras geradas por um mesmo protocolo amostral. É fato que os erros amostrais variam em função dos métodos de amostragem, preparação e análise empregados, mas todos compartilham uma mesma ordem de magnitude, o que torna pouco justificável classificar parte dos dados como *hard data*, cujo o erro é ignorado, e um ou mais grupos de *soft data*. O erro associado a uma amostra não é produto apenas do protocolo empregado. Mudanças não documentadas do processo e na propriedade do material analisado, variações na qualidade dos insumos usados nas análises, entre vários, também controlam a qualidade dos dados (RAMSEY; ELLISON, 2007).

Os métodos de simulação geoestatística (JOURNEL, 1974; ISAAKS, 1990) são usados para caracterizar a dispersão espacial de um fenômeno real e a incerteza associada a esse modelo. Nas práticas atuais, é assumido que o erro amostral é indiretamente considerado pela krigagem através do aumento do efeito pepita no variograma empregado. Na presença de erros, as práticas correntes de simulação têm dois problemas:

- simulações condicionais honram os valores *hard data*, considerando o intervalo de incerteza como nulo nas posições amostradas, subestimando a incerteza local;

[1] Os termos *hard data* e *soft data,* empregados ao longo do trabalho, têm o mesmo significado de dados primários e secundários. Tais termos são utilizados de forma intercambiáveis na literatura.

● histogramas, variogramas e a estrutura de correlação medida através das observações combinam o fenômeno de interesse e o erro de amostragem. A reprodução da componente de erro pela simulação leva à superestimava da variabilidade do fenômeno simulado. Portanto, realizações condicionadas às estatísticas dos dados não são equiprováveis ao fenômeno real, pois não reproduzem simultaneamente os valores do fenômeno real nos locais amostrados, o covariograma e a distribuição do fenômeno real de interesse.

Hard data não podem ser obtidos por amostragem, mas valores equiprováveis a eles podem ser simulados a partir de observações com erro. Bancos de dados compostos apenas por *hard data* podem ser gerados ao substituir as amostras iniciais (*soft data*) por simulações do fenômeno real. Cada novo banco é, então, utilizado para simular o modelo em todo o domínio, condicionado a honrar os *hard data* e as estatísticas inferidas do fenômeno real. Nesse caso, o resultado são modelos realmente equiprováveis ao fenômeno real.

Nesta tese, serão discutidas e desenvolvidas todas as etapas e métodos necessários para lidar com problemas geoestatísticos na presença de erros de amostragem em todas as observações, considerando o seu caso mais complexo. Portanto, o modelo proposto leva em conta problemas no seguinte contexto:

● fenômenos multivariados – os valores e probabilidades condicionais de interesse não são apenas os individuais de cada propriedade, mas também a estrutura de correlação entre elas;

● observações afetadas por erros heterogêneos – o comportamento do erro associado a diferentes amostras é heterogêneo, e, portanto, agrupá-los em subgrupos causa perdas;

● o comportamento do erro associado a cada valor individual pode ser estimado – o modelo do comportamento de erros considera a possibilidade de que os erros fossem dependentes do teor, e, que no caso multivariado os erros possam ser correlacionados entre diferentes variáveis.

Tal complexidade garante a generalidade do método, possibilitando seu uso para diferentes depósitos ou problemas geoestatísticos sem simplificações ou sendo necessário ignorar características relevantes do depósito e das observações disponíveis.

1.1 META

Hoje, é completamente plausível a ideia de se ter uma estimativa do erro associado à cada medição disponível em um banco de dados. O entendimento do comportamento do erro e o emprego de programas de monitoramento sistemático dos seus valores se tornaram na indústria mineral algo comum, e quase obrigatório, ao longo das últimas décadas. Este contexto é muito diferente daquele do estabelecimento da geoestatística (KRIGE, 1951; MATHERON, 1963), em que rotinas de monitoramento do erro total das amostras geradas eram inexistentes e o entendimento das suas origens e formas de prever seu comportamento ainda viriam a ser desenvolvidos, como pelo trabalho desenvolvido por Gy (1968).

As estimativas e a incerteza medida em um modelo não devem considerar apenas a distância entre as observações e sua variabilidade local, mas também o erro associado aos dados. Essa afirmação e suas implicações são analisadas nos métodos de simulação geoestatística, os quais são utilizados para a geração de realizações geralmente assumidas como equiprováveis ao fenômeno de interesse. Temos a seguinte passagem de Journel e Huijbreigts (1978), um dos primeiros trabalhos a abordar o tema:

> A regionalized variable z(x) is interpreted as one realization of a certain random function Z(x). This RF - more precisely this class of RF's - is characterized by a distribution function and a covariance or variogram model. The idea of simulations consists in drawing other realizations Zs(x) from this class of RF's. (JOURNEL; HUIJBREIGTS, 1978, p. 491).

Tal definição é base para métodos de simulação posteriormente desenvolvidos. O ponto de atenção é que quando as observações são afetadas por erros, a distribuição e o variograma ajustados às observações não são iguais aos da função aleatória a ser simulada. As alternativas propostas simplificam o problema ao assumir que parte dos dados é isenta de erros, e, então, a distribuição dos valores e variogramas ajustados a eles são representativos do fenômeno real, o que é incorreto. No entanto, novamente é afirmado que *hard data* não podem ser obtidos por amostragem. E, portanto, a variância dos dados é a combinação da variância do fenômeno real e da variância dos erros de amostragem e medição. Por consequência, realizações cujo os histogramas, variogramas e a estrutura de

correlação a quais são condicionadas são calculados a partir de observações, não são equiprováveis ao fenômeno real.

Dada as rotinas atuais de estimativa e simulação da indústria mineral que ignoram a presença e/ou complexidades dos erros associados aos dados, esta tese é baseada na seguinte afirmação, tomada como hipótese de pesquisa:

Hipótese de pesquisa:
Amostras isentas de erros são inexequíveis. Ao ignorar a presença do erro associado aos dados, as rotinas atuais de simulação levam a realizações que não são equiprováveis ao fenômeno real.

A resposta da hipótese leva a temas inexplorados na literatura geoestatística, que são métodos de simulação aplicáveis para casos em que parte ou todo o banco de dados não são assumidos como isentos de erro.

1.2 OBJETIVOS

O objetivo desta tese é investigar métodos de simulação capazes de gerar modelos que lidem corretamente com os erros individuais associados a todas as observações disponíveis, levando a possibilidade de se obter realizações que realmente sejam equiprováveis ao fenômeno real e que meçam corretamente a incerteza entre o modelo e o fenômeno real.

A meta e objetivos são alcançados através dos seguintes objetivos específicos:

- definir um modelo do comportamento do erro amostral que seja válido tanto o caso univariado quanto multivariado;
- desenvolver alternativas para estimar os parâmetros do modelo de erro;
- desenvolver alternativas para transformar erros amostrais em unidades Gaussianas;
- desenvolver formas para estimar a correlação espacial, estrutura de correlação e a distribuição do fenômeno real através de observações com erro;
- desenvolver um sistema de krigagem que considere o erro individual de cada observação na definição dos seus pesos;
- desenvolver métodos de simulação estocástica condicionada aos parâmetros do fenômeno real e à incerteza de cada observação.

Como delimitação do escopo, esta tese lida com os tipos de dados e erros associados à indústria mineral. As observações são análises de pequenas frações vindas de amostras de rocha ou solo que passam por sucessivas etapas de redução granulométrica e

quarteamento do volume original. Essa realidade difere dos dados obtidos em reservatórios convencionais de petróleo, onde as medições em poços (*hard data*) vêm de uma média de dezenas ou centenas de medições individuais; e de dados sísmicos (*soft data*), medições indiretas que acumulam diversos erros de medição e posicionamento. Nesse contexto, assumir o primeiro como *hard data* é plausível. Ainda assim, todo o método desenvolvido pode ser estendido para outras situações, desde que observadas suas especificidades.

Toda a metodologia aqui desenvolvida, assim como as equações e os sistemas de krigagem propostos, são aplicáveis para problemas da área de reservatórios de hidrocarbonetos, apesar da nomenclatura e exemplos focados em problemas de mineração. Caso o interesse do leitor seja essa aplicação, o termo "teor" que é extensivamente usado, pode ser entendido como a propriedade de interesse a ser simulada, como porosidade ou permeabilidade. O erro de medição associado aos dados sísmicos naturalmente não é abrangido pelos métodos da Teoria da Amostragem (Gy, 1982), sendo isso discutido na seção "trabalhos futuros".

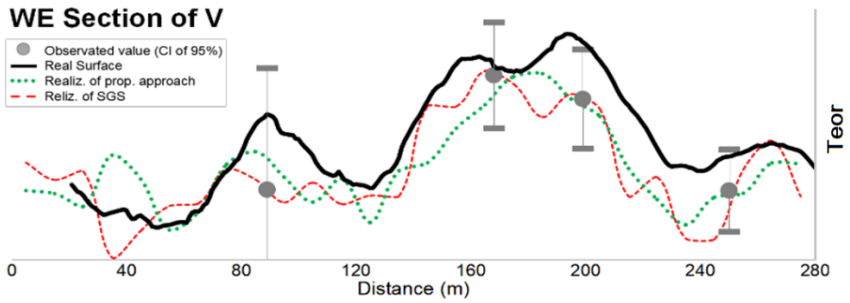
1.3 METODOLOGIA

A solução para lidar com amostras e as suas incertezas associadas é um problema adequado à modelagem estocástica, em que possíveis realizações do fenômeno real são simuladas. O método proposto é composto de duas etapas: primeiro, os valores amostrais são substituídos um a um por realizações equiprováveis ao fenômeno de interesse. Em seguida, todo o modelo (nós amostrados ou não) é simulado utilizando os bancos de dados simulados. Em ambos os casos, as realizações são condicionadas a honrar a distribuição e variograma inferidas do fenômeno real. Como resultado da remoção da componente de erro, o método proposto tem realizações com menor variância global e maior continuidade espacial do que as abordagens convencionais (JOURNEL, 1974; ISAAKS, 1990).

A Figura 1 ilustra as principais diferenças entre a metodologia proposta (linha verde), a qual reproduz o histograma e o variograma inferidos do fenômeno real; e o método convencional (linha vermelha), que reproduz os valores, histograma e variograma ajustados às observações assumidas como *hard data*. No exemplo, observações (círculos cinzas) do fenômeno real (linha preta) são obtidas em diferentes posições.

Uma primeira característica é que o método proposto não honra as observações, mas sim o espaço de incerteza dessas observações (faixas verticais) nessas posições. Outra diferença é que esse método apresenta uma menor variabilidade entre locais amostrados, já que é condicionado a honrar variogramas e histogramas sem a componente de erro.

Figura 1 – Comparação entre realizações do método proposto e convencional em relação ao fenômeno real



Fenômeno real (linha preta) simulado pela abordagem proposta (linha verde) e método convencional (linha vermelha). O método convencional honra os valores amostrais (círculos cinzentos), histograma e variograma dos dados. O método proposto honra as estatísticas inferidas do fenômeno real e o espaço de incerteza das observações (linha cinza vertical).

De forma detalhada, as etapas da metodologia são:

i. Estimando os erros individuais de cada amostra: a incerteza associada a cada observação é definida em subgrupos que têm a mesma relação entre teor e erro. Essa relação que caracteriza o subgrupo é uma função das características do material amostrado, protocolo de amostragem etc. A relação modelada para cada subgrupo é usada para estimar a incerteza dessas observações.

Entre os métodos apresentados, o mais geral emprega diversos critérios para delimitação de subgrupos. Amostras são agrupadas em função do seu tipo de rocha, protocolos de amostragem utilizados e o comportamento de duplicatas. Tais dados são considerados em etapas sucessivas testes de significância estatística e análise gráfica para delimitação de possíveis do subgrupo.

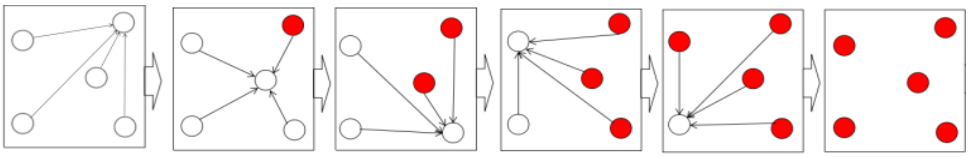
ii. Transformando os erros para uma distribuição Gaussiana: o uso de métodos geoestatísticos que se baseiam nas propriedades da distribuição muitiGaussiana torna necessário transformar os dados iniciais para tal distribuição. Ao lidar com erros

associados a tais valores, também é necessário transformá-los para as mesmas unidades. São apresentadas soluções analíticas e computacionais.

iii. Gerando realizações do fenômeno real: a simulação do fenômeno real, a partir de observações com erros, tem três passos: são inferidos os parâmetros do fenômeno real de interesse, simulados bancos de dados de *hard data* e, então, as realizações no domínio de interesse são simuladas. Cada realização é condicionada a honrar os parâmetros inferidos do fenômeno real e um banco de dados diferente.

a. **Inferindo os parâmetros reais:** os modelos de covariogramas e histogramas são estimados através da remoção da componente de erro dos modelos ajustados aos dados. Aqui surge a maior vantagem do método de estimativa apresentado: a estrutura de covariância espacial pode ser estimada diretamente do covariograma ajustado às observações que apenas requer o conhecimento dos erros associados à tais observações.
b. **Simulando os bancos de dados:** As observações que compõem o banco de dados inicial são combinações dos valores do erro amostral e do fenômeno real. Sendo o comportamento do erro conhecido, o fenômeno real pode ser simulado nos pontos amostrados. O algoritmo visita cada posição, uma por uma, para substituir os valores amostrados por realizações do *hard data* (Figura 1). A probabilidade local é estimada através de *full co-kriging* (COK; MARECHAL, 1970; JOURNEL; HUIJBREIGTS, 1978) utilizando os dados iniciais, colocalizados ou não, e os *hard data* já simulados na vizinhança

Figura 2 - Esquema da substituição sequencial de observações por *hard data*



Esquema da substituição sequencial dos valores iniciais (círculos vazios) por realizações do fenômeno real (círculos vermelhos).

c. **Simulando o modelo completo:** Os bancos de dados simulados e os parâmetros inferidos são considerados na simulação de todo o modelo. É recomendado que seja simulado um novo banco de dados para cada simulação do modelo completo.

iv. Pós-processamento das realizações: O modelo de incerteza simulado é, geralmente, pós-processado para gerar um segundo produto. São exemplos de funções que podem ser aplicadas ao conjunto de nós simulados: *e-type*; cálculo da variância da distribuição condicional; a probabilidade de exceder um dado limite e o valor médio acima (ou abaixo) desse limiar etc.

A metodologia acima é desenvolvida para o caso univariado, ignorando a associação entre variáveis. No entanto, parte dos problemas geoestatísticos é multivariada, o que justifica estender o fluxo de trabalho para a situação em que a estrutura de covariância cruzada entre diferentes fenômenos reais se torna relevante. Abaixo as principais mudanças na metodologia:

a. **Modelando a estrutura de correlação entre dois fenômenos:** no caso multivariado, os erros entre duas ou mais variáveis podem ser correlacionados. A inferência da correlação real através de observações deve considerar a proporção dos erros que sejam dependes ou independentes ao teor, assim como se os erros são correlacionados ou não entre as variáveis.

b. **Gerando realizações do fenômeno real:** A co-krigagem intrínseca colocalizada (ICCK; BABAK; DEUTSCH, 2009a) substitui a co-krigagem (COK; MARECHAL,

1970) no caso multivariado para estimar as probabilidades locais na simulação. Essa mudança é necessária, pois sistema de krigagem se torna muito grande e pode se tornar instável ao lidar com a covariância cruzada entre os valores iniciais colocalizados e na vizinhança e os valores já simulados de diferentes variáveis.

Nos apêndices, são discutidas mudanças na metodologia para lidar com variáveis categóricas, além da possibilidade de utilizar observações com viés.

1.4 INTEGRAÇÃO DE ARTIGOS

A presente tese se baseia, principalmente, no conteúdo de diferentes produções científicas de autoria do pesquisador deste trabalho - sendo três artigos publicados e dois em processo de publicação, todos em periódicos internacionais. Apesar da redundância entre os artigos e as seções da tese, foi decidido apresentá-los das duas formas para melhorar a consistência entre terminologias e notações dos diferentes trabalhos.

O primeiro artigo (anexo 1), “Selecting The Maximum Acceptable Error In Data Minimising Financial Losses” (SILVA; COSTA, 2016a), apresenta o arcabouço conceitual do problema de otimização de posicionamento à relevância de levar em conta na otimização a definição de qual o melhor protocolo amostral a ser usado em cada posição. A relação entre o custo da aquisição de novas amostras e o retorno através da melhora das decisões tomadas com tal modelo, como na classificação de blocos de minério e estéril, deve ser levada em conta. Ganhos financeiros e a eficiência do programa de sondagem podem ser obtidos não apenas definido a melhor posição para se adquirir amostras, mas também qual o melhor protocolo amostral para se empregar em casa posição.

O primeiro trabalho levanta a relevância de se considerar que, no mundo real, bases de dados podem ser compostas não de um ou dois subgrupos de amostras, mas de diversos subgrupos. Para lidar com eles, é necessário estimar o erro associado a cada observação. O anexo 2 apresenta o artigo “Using QC data to estimate the individual precision of original samples: a duplicates-based approach” (SILVA; COSTA, 2018), em que se propõe um método para estimar o erro associado a cada observação através de amostras de QAQC.

Sendo possível estimar o erro individual associado a cada observação, o trabalho “Geostatistics in the presence of sampling error: Simulating the underlying error-free true

process from observations with error" apresenta uma série de métodos estatísticos e geoestatísticos que possibilite lidar com os diversos problemas de simular modelos univariados na ausência de dados isentos de erro. Tal trabalho foi submetido à revista "Mathematical Geosciences" e está em fase de revisão. A seção "Simulando o Processo Real a Partir de Observações: caso Univariado" é baseada, principalmente, neste artigo.

No anexo 3, o artigo "Sampling error correlated among observations: Origin, impacts and solutions" (SILVA; COSTA, 2019) apresenta o arcabouço conceitual e o relaciona ao contexto de amostragem de materiais geológicos em que ocorre erros compartilhados. Tal artigo é relevante para possibilitar que a extensão dos métodos estatísticos e geoestatísticos apresentados acima sejam adaptadas ao caso multivariado. Um artigo baseado na seção "Simulação na presença de erros: caso multivariado" está em confecção. Além dos trabalhos acima, diversos artigos não revisados pelos pares foram publicados[2].

1.5 ESTRUTURA DA TESE

A tese desenvolve rotinas de simulação que lidam de forma adequada quando há a disponibilidade apenas de dados afetados por erros (os chamados *soft data*). Nesta tese, foram desenvolvidas todas as etapas necessárias para a modelagem sob essa condição tanto para o caso univariado quanto no caso multivariado.

O capítulo 2 apresentada um levantamento bibliográfico que discute os diferentes métodos geoestatísticos que se propõem a lidar com os erros associados aos dados.

O capítulo 3 revisa o comportamento do erro amostral e define um modelo geral para ser empregado no caso univariado. Diversas alternativas para estimar os parâmetros de tal

---

- [2] SILVA, V.M.; COSTA, J.F.C.L.; DEUTSCH, C.V. A Short Note On The Relationship Between Relative Original Units Error And Absolute Normal Score Error. *In*: CCG 21TH ANNUAL MEETING, 2019, Edmonton. CCG 21th Annual Meeting, 2019.
- SILVA, V.M.; COSTA, J.F.C.L.; DEUTSCH, C.V. The Model of Additive Rescaled Error (MARE). *In:* CCG 21TH ANNUAL MEETING, 2019, Edmonton. CCG 21th Annual Meeting, 2019.
- SILVA, V.M.; COSTA, J.F.C.L.; DEUTSCH, C.V. Inference of True Data Parameters Given Data with Error. *In:* CCG 20ST ANNUAL MEETING, 2018, Edmonton. CCG 21st Annual Meeting, 2018.
- SILVA, V.M.; COSTA, J.F.C.L.; DEUTSCH, C.V. Imputation of True Data Values Given Data with Error. *In:* CCG 20TH ANNUAL MEETING, 2018, Edmonton. CCG 20th Annual Meeting, 2018.

modelo são discutidas, tais como: o uso do efeito pepita, a teoria da amostragem ou duplicatas geradas por programas de QA/QC. Após definido o modelo de erro, são apresentadas formas de transformar tais valores para valores Gaussianos e, então, é definido um modelo de krigagem capaz de estimar a distribuição local de probabilidades (cdf) quando todos os dados são afetados por erros conhecidos. A krigagem se baseia em um modelo linear de regionalização construído em função dos erros associados a cada par de nós usado em seu cálculo, sendo eles amostrados ou não. Por fim, um fluxo de simulação capaz de lidar com os erros de cada medição é desenvolvido. Todo o desenvolvimento é ilustrado com diversos exemplos, como forma de elucidar a relevância de cada etapa.

O capítulo 4 estende toda a teoria para o caso multivariado. Primeiro, o modelo é adaptado de forma a considerar que erros podem ser correlacionados entre diferentes variáveis. O impacto de ignorar esse tipo de erro é demonstrado analiticamente e através de exemplos. São desenvolvidas formas de estimar a distribuição, o covariograma direto e cruzado e a estrutura de covariância entre diferentes fenômenos reais através de observações. Um sistema de co-krigagem mais geral, que lida com todos os tipos de erros, é desenvolvido e seu uso para definir a CDF necessária para simulação dos fenômenos de interesse é apresentado.

O capitulo 5 resume as principais contribuições da tese, analisa as soluções dos objetivos inicialmente propostos e apresenta sugestões para trabalhos futuros.

A tese também contém dois apêndices. No primeiro, são discutidas mudanças no fluxo proposto para lidar com observações enviesadas, seja o viés conhecido ou não. No segundo apêndice, alternativas e ideias de formas para definir as componentes de erro compartilhadas são discutidas.

Nos três anexos da tese, são apresentados os artigos publicados ao longo do desenvolvimento da pesquisa. O anexo 1 apresenta o artigo "*Using QC data to estimate the individual precision of original samples: a duplicates-based approach*" (SILVA; COSTA, 2018), em que se propõe um método para estimar o erro associado a cada observação através de amostras de QAQC. A seção 3.2.4 Definindo subgrupos de erro através de amostras duplicatas) se baseia diretamente neste trabalho. No Anexo 2, o artigo "Sampling error correlated among observations: Origin, impacts and solutions" (SILVA; COSTA, 2019) apresenta o arcabouço conceitual e o relaciona ao contexto de amostragem de materiais

geológicos onde ocorre erros compartilhados. A seção 4.1 ESTENDENDO O MODELO DE ERROS PARA O CASO MULTIVARIADO se baseia diretamente neste trabalho. Por fim, o artigo "*Selecting The Maximum Acceptable Error In Data Minimising Financial Losses*" (SILVA; COSTA, 2016a) demonstra a relevância e expectativa de ganhos de uma das principais aplicações diretas da tese, que é o de ser possível, em métodos de otimização de amostragem, considerar não apenas o posicionamento amostral, mas também qual o melhor protocolo a ser usado naquela posição em específico. Isso só é possível se a solução usada for capaz de considerar a contribuição individual de cada amostra no espaço de incerteza simulado, assim como o de considerar simultaneamente qualquer quantidade de dados com diferentes erros.

2 REVISÃO DA LITERATURA

Nesta seção, são apresentados diferentes métodos geoestatísticos que se propõem a lidar com os erros associados aos dados. Não é escopo da tese ou desta seção esgotar o tema de simulação ou erro amostral. Para conceitos básicos e métodos já estabelecidos, como krigagem ordinária (MATHERON, 1963), simulação sequencial gaussiana (ISAAKS, 1990) e simulação sequencial direta (SOARES, 2001) é recomendado recorrer a artigos sobre o tema e livros textos, tais como: Goovaerts (1997), Isaaks e Srivastava (1989) e Wackernagel (2003). Referencias adicionais são apresentadas quando necessário ao longo do texto. A revisão é agrupada em três subseções: (i) o erro amostral; (ii) estimativa na presença de erros; e (iii) simulação na presença de erros.

2.1 O ERRO AMOSTRAL

O erro amostral é intrínseco ao processo de amostragem de materiais heterogêneos, como rochas, mesmo quando o protocolo amostral empregado é teoricamente correto e perfeitamente executado (GY, 1982). Apesar não ser possível extinguir o erro, ele é monitorado e minimizado através dos chamados “programas de controle e garantia da qualidade”, daqui em diante tratados pelo acrônimo em inglês QA/QC. Mesmo quando seguidas as melhores práticas de amostragem e de controle de qualidade, o erro total medido por duplicatas tem um coeficiente de variação entre 15% e 40% (ABZALOV, 2008) e, portanto, o erro é sempre relevante e deve ser considerado nos fluxos geoestatísticos (SILVA, 2015; SILVA; COSTA, 2016a; 2016b). Ainda assim, é prática comum desconsiderar o erro associado aos dados aprovados pelo programa de QA/QC. Sabe-se que bancos de dados reais podem combinar dados com diferentes tipos e magnitudes de erros. Isso ocorre mesmo que todos dados tenham sido gerados ao longo do tempo pelo mesmo protocolo. A teoria de amostragem (GY, 1982; PITARD, 1993) ou TOS, fornece à incerteza amostral uma definição cientifica e um esclarecimento conceitual. A TOS define os erros de amostragem como produtos do não atendimento da equiprobabilidade amostral de cada partícula em materiais heterogêneos no processo de coleta, divisão mássica e análise laboratorial.

Os erros amostrais não são produtos apenas do tipo de material ou do protocolo utilizado. Mesmo em processos bem documentados e monitorados haverá variações devido a ambiguidades, adaptações e mudanças não documentadas feitas ao protocolo (RAMSEY;

ELLISON, 2007). Essas diferentes origens do erro adicionam complexidade ao lidar com a modelagem e estimativa do erro de bases de dados reais. Entendermos o comportamento do erro amostral é necessário para o integrá-lo aos métodos geoestatísticos.

Primeiro, o erro total é a combinação de diversas fontes que se combinam ao longo de todo o processo de coleta, redução granulométrica, divisão mássica e análise. Segundo o teorema do limite central, o somatório desses processos independentes leva a uma distribuição gaussiana. Thompson e Howarth (1976) alertam que a gaussianiedade da distribuição do erro sempre deve ser validada, pois há risco de se obter conclusões significativamente enviesadas nas raras distribuições de erro significantemente divergentes dessa hipótese.

Um segundo ponto a se investigar é a possível associação entre teor e erro amostral. Estudos demonstram ser inapropriado utilizar o erro médio para valores distribuídos em faixas maiores que uma ordem de magnitude em relação à média. Nesses casos, pelo menos parte do erro tem uma relação linear com o teor (THOMPSON; HOWARTH, 1973, 1976, 1978; HOWARTH; THOMPSON, 1976). Uma componente independente do teor geralmente está presente. Há também fontes de variação que podem fazer o comportamento não ser linear ou se deparara com comportamentos específicos, como erros não-gaussianos (STANLEY; LAWIE, 2008), variações com alto efeito pepita (STANLEY, 2006) ou quando o comportamento do erro de amostragem e de análise são diferentes em relação ao teor (FRANCOIS-BONGARCON, 1998).

2.2 ESTIMATIVA NA PRESENÇA DE ERROS

A krigagem em sua versão inicial (MATHERON, 1963) é apresentada como um interpolador exato, considerando que a melhor estimativa no local onde há um ponto amostrado é o valor que ali se encontra. Tal solução é válida ao se assumir o efeito pepita como resultado de variações geológicas não capturadas pelo espaçamento ou suporte amostral. No entanto, o efeito pepita C(**0**) pode ser decomposto em erro de medição C_{se} e em variância de microescala C_{me}, e nesse caso, proporção das componentes influenciam a forma como a krigagem deveria se comportar (CRESSIE, 1993; PITARD, 1993; FRANCOIS-BONGARÇON, 2004). Apesar da SK e OK serem geralmente tratados como interpoladores exatos, a exatidão, ou não, é produto da interpretação da origem do efeito pepita. Na

prática, C(0) é geralmente obtido ao interpolar os primeiros passos do variograma até a origem, e todo esse valor atribuído a variações de micro-escala C(**0**) = C_{me}. Em caso de o efeito pepita ser composto apenas por uma das componentes, teremos:

1. Todo o efeito pepita é causado por variações de microescala, C(**0**) = C_{me} e C_{se} = 0, tornando o método um estimador exato.

2. Na presença de erros de medição, (C_{se} > 0), resolve-se os pesos de krigagem λ considerando que a variância do valor de um ponto e ele mesmo não é mais zero.

Como limitação, mesmo que o erro individual de cada amostra seja conhecido, o sistema de krigagem irá considerar apenas o valor médio C_{se} do modelo variográfico. Como solução, a Co-krigagem (CoK; MARECHAL, 1970) utiliza os variogramas diretos e cruzados de grupos de amostras com erros amostrais e suportes diferentes, não havendo do ponto de vista teórico diferença para a krigagem (JOURNEL; HUIJBREIGTS, 1978). O aumento do número de variogramas diretos e cruzados requeridos e das restrições para definir o modelo linear de corregionalização faz com que, na prática, os dados sejam agrupados em no máximo três grupos devido a quantidade de variogramas diretos e cruzados a serem modelados se tornar impeditiva após esse ponto. Quão maior for a heterogeneidade em cada subgrupo (diferentes erros médios, diferentes suportes amostrais etc.) maiores serão as perdas de acurácia do modelo.

Quando as amostras têm seus erros de amostragem conhecidos, tenham um mesmo suporte e sejam independentes do teor, tal erro pode ser considerado na Krigagem com Variância do Erro de Medição KVME (DELHOMME; 1976; WACKERNAGEL, 2003). O KVME se baseia no fato em que o covariograma não é afetado por erros independentes do teor, Cov{$z_{k,obs}$(**u**), $z_{k,real}$(**u**)} = Cov{$z_{k,real}$(**u**), $z_{k,real}$(**u**)} quando **h** > 0 e Cov{ε(**u**), z_{real}(**u**)} = 0. A diagonal da matriz do KVME (**h** = 0), que corresponde a covariância entre amostras, é preenchida pela variância do erro associado a cada amostra pois a covariância de um ponto com ele mesmo é sua variância e como resultando, quão maior for o ε(**u**) associado a cada medição do z_{real}(**u**), menor o peso atribuído a esse valor. No entanto, como limitação do método, a condição de independência entre erro e teor raramente é satisfeita em problemas reais (HOWARTH; THOMPSON, 1976; THOMPSON; HOWARTH, 1973, 1976, 1978).

A krigagem dos indicadores com *soft data* sIK (JOURNEL, 1986; ZHU; JOURNEL, 1993) se baseia na krigagem dos Indicadores (IK), mas considera a incerteza do valor lido no ponto amostral na codificação do indicador I(**u**;Z) = *Prob*{Z(**u**) < z}. O método requer o conhecimento do erro associado a cada observação e, ao empregar formalismo de indicadores para caracterizar variáveis continuas, traz consigo problemas de relação de ordem e perda de resolução dos valores locais. Outro ponto é que o modelo de incerteza fornecido pelo IK se aplica apenas ao suporte das amostras, que, geralmente, é muito menor do que o tamanho dos volumes a serem estimados, tornando necessário corrigir as probabilidades *a posteriori* para tal suporte. Por fim, ao discretizar os dados em mais classes, o uso de diferentes variogramas em cada classe pode causar problemas de relação de ordem entre as probabilidades estimadas para as várias classes (GOOVAERTS, 1997).

Existem também alternativas que medem o impacto da variância do erro através do uso de geradores pseudoaleatórios para a adição de novos erros. O erro sintético de cada amostra e seu impacto nas estimativas é medido através do uso de diferentes bancos de dados com crescentes níveis de erros adicionados. O método foi empregado para comparar o desempenho de diferentes métodos de krigagem para integrar simultaneamente dados com maior e menor qualidade (MAGRI; ORTIZ, 2000; ARAÚJO; COSTA, 2015), assim como adicionando erros sintéticos, tanto aditivos quanto multiplicativos, para analisar como estes afetam as estimativas (EMERY *et al.*, 2005). A geração de bancos de dados com diferentes níveis e erros também foi empregada para modelar a relação entre custos e o retorno de diferentes protocolos amostrais (SILVA; COSTA, 2016a), assim como a relação entre gastos e retorno financeiro do emprego de furos de desmonte ou de circulação reversa em rotinas de curto prazo (ORTIZ *et al.,* 2012). Apesar de honrar local e globalmente a média dos dados, as bases de dados gerados pela adição de erros têm uma degradação da correlação espacial e um aumento da variância maior do que teriam dados reais com o mesmo valor de erro.

Os resultados de Emery (2005) e Cornah; Machaka (2015) sintetizam a relação das estimativas que empregam *hard data* e *soft data* e concluem que que dados com alta incerteza não devem ser descartados. Através da análise para diferentes depósitos com diferentes proporções entre dados de maior ou menor erro associado, conclui-se que a quantidade de dados compensa a baixa qualidade das observações quando o método empregado é capaz de levar em consideração as especificidades dos dados. Entre os

diversos métodos acima listados, tais como OK, CoK, sIK, KVME, a escolha de qual seria o mais adequado deveria considerar a magnitude do erro associado a cada medição, se os dados são afetados por viés e se o erro individual de cada dado pode ser estimado.

2.3 SIMULAÇÃO NA PRESENÇA DE ERROS

Na literatura, há trabalhos de simulação geoestatística que propõem formas de considerar o erro amostral ao quantificar o espaço de incerteza do modelo. As primeiras tentativas de lidar com esse problema são divididas por Soares *et al.* (2016) naqueles que obtêm a CCDF local através de sIK (SRIVASTAVA, 1992; FROIDEVOUX, 1993) e naqueles que obtêm a CCDF através da co-krigagem dos indicadores com *soft data* (JOURNEL, 1986; ZHU; JOURNEL, 1993). Ambos compartilham limitações relacionadas a dificuldade do emprego do formalismo de indicadores para lidar com variáveis continuas, principalmente, o modelo de estimativa de covariância (GOOVAERTS, 1997). A literatura atual sobre simulação estocástica na presença de erros pode ser separada em dois grupos:

- aqueles que consideram o espaço de incerteza associado a cada dado no cálculo de suas probabilidades condicionais;
- métodos com duas etapas – uma etapa onde são simuladas realizações dos bancos de dados e, então, empregadas como *input* para simular todo o modelo, nos demais nós, amostrados ou não.

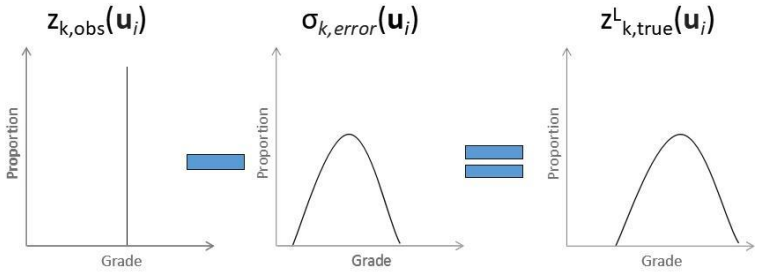
Marcotte (1995) examina o uso de krigagem fatorial (SANDJIVY, 1984) nas etapas de pré ou pós-processamento para filtrar da simulação condicional a componente do erro de medição. A filtragem pré-simulação é rejeitada porque é ineficiente e o variograma esperado não é reproduzido. As filtragens pós-simulação são tratáveis e eficientes. As realizações podem ser geradas pelo método de bandas rotativas (JOURNEL; HUIJBREGTS, 1978) ou outros como sGs e simulação sequencial direta (DDS).

O método de campo de probabilidade *p-Field* (SRIVASTAVA, 1992; FROIDEVOUX, 1993) emprega sIK para incorporar os erros de medição. O valor krigado e sua variância são, então, usados para interpolar no espaço essas distribuições de probabilidade local. Uma vez que a incerteza decorrente de erros de medição e interpolação espacial foi modelada, a simulação de p-Field leva a uma rápida agregação das CCDF pontuais para derivar a probabilidade de

um local exceder um limiar específico no suporte desejado (SAITO; GOOVAERTS, 2002) através da discretização do volume em pontos simulados (JOURNEL; HUIJBREGTS, 1978). As limitações do *p-Field* incluem a falta de base teórica para simulações condicionais (JOURNEL, 1995). Além disso, as realizações apresentam dois artefatos severos quando os nós simulados estão próximos à *hard data*, os valores condicionais se mantêm como o valor mínimo e máximo entre as simulações e as realizações são significantemente mais continuas que os dados condicionais (PYRCZ; DEUTSCH, 2001).

O segundo conjunto de métodos considera que, apesar de *hard data (livre de erro)* não poderem ser obtidos, múltiplas realizações equiprováveis ao fenômeno de interesse podem ser simuladas nos locais onde esses foram amostrados (Figura 3).

Figura 3 - Simulando $z_{k,true}(\mathbf{u})$ através de $z_{k,obs}(\mathbf{u})$ e $\sigma^2_{k,obs}(\mathbf{u})$



Na presença de uma observação $z_{k,obs}(\mathbf{u})$, podemos simular realizações (L) $z^L_{k,real}(\mathbf{u})$ do fenômeno real condicionado ao valor observado, distribuição com erros $\sigma^2_{k,erro}(\mathbf{u})$ e outras probabilidades condicionais.

Entre os métodos que se propõe a simular possíveis variações dos dados disponíveis, Cuba *et al.* (2008; 2012) um amostrador de Gibbs é empregado para obter valores que substituam as observações disponíveis, respeitando em locais amostrados o espaço de incerteza dessa observação (dada pelo valor conhecido e seu erro amostral) e globalmente, honrar o variograma e a distribuição dos dados. O método pode apresentar baixa eficiência computacional devido à grande quantidade de interações necessárias para que os resultados convirjam para uma realização aceitável através de tentativa-e-erro. Além disso, ao gerar realizações do banco de dado inicial sem remover a componente de erro associada aos dados, leva a realizações equiprováveis aos valores amostrais, mas não equiprováveis ao fenômeno real.

Soares *et al.* (2016) propõem substituir os dados iniciais por realizações que são equiprováveis aos dados considerados como medições sem erro. As simulações são realizadas por DDS, que dispensa qualquer transformação das observações para uma

segunda distribuição (SOARES, 2001). A simulação usa os *hard data* disponíveis e os dados previamente simulados. O método não considera explicitamente o erro de amostragem, assumindo que na estimativa por SK em cada posição utilizando o variograma sem erro resulta em uma estimativa correta da distribuição local. O erro de medição médio é considerado através da sua componente no efeito pepita do variograma utilizado. A limitação do método é que ao lidar com erros que variam espacialmente, tal variação não é diretamente considerada na estimativa da CCDF local. Como forma de considerar mais de uma população de erros, Araújo (2019), em uma primeira etapa, substitui os valores assumidos como *soft data* por valores equiprováveis ao fenômeno real e condicionados aos parâmetros ajustados aos dados assumidos como *hard data*. O método compartilha a limitação de ser necessário agrupar os dados em um ou no máximo dois grupos de *soft data* como forma de viabilizar a modelagem do modelo linear de corregionalização. Além disso, é necessário assumir parte dos dados como *hard data*, o que é irrealista, como já discutido.

O contexto apresentado e o método proposto podem ser vistos como um problema de imputação em que o *hard data* é ausente em todas as posições, onde uma simulação pontual é feita no local dos dados. A noção básica de imputação em um contexto de modelagem geoestatística é a de construir distribuições condicionais representativas nos locais a se simular realizações da variável de interesse faltante. No caso multivariado, seriam necessárias distribuições para cada variável. Simular com base nessas distribuições condicionais gerará várias realizações dos dados a serem usados para condicionamento de simulação geoestatística.

Ao considerar a imputação de valores de uma variável regionalizada, a distribuição condicional de possíveis valores de *hard data* pode ser inferida com base nos (i) valores de observações disponíveis da mesma variável e (ii) os valores de observações colocalizadas e *soft data* correlacionados ao processo de interesse. A construção de distribuições condicionais a partir dessas duas fontes é bem definida sob a hipótese multiGaussiana com métodos como CoK ou atualização bayesiana (BARNNET; DEUTSCH, 2012, 2015; REN, 2007). Enquanto o primeiro exige definir o modelo linear de corregionalização, a atualização bayesiana requer apenas o coeficiente de correlação entre as observações e o fenômeno real.

Barnett e Deutsch (2015) propuseram duas metodologias de múltiplas imputações para dados faltantes através de atualização bayesiana: a abordagem não-paramétrica calcula a distribuição conjunta a partir dos gráficos de dispersão de um subconjunto isotópico entre *hard* e *soft data*. O método paramétrico assume que a distribuição conjunta é MultiGaussiana, sendo necessário dados conhecer o coeficiente de correlação entre as variáveis, o que não é diretamente disponível na presença de erros amostrais.

A limitação dos métodos de imputação, é que, n contexto da pesquisa, temos um caso em que não existe *hard data* e, portanto, as estatísticas multivariadas entre o fenômeno real e as observações não estão disponíveis de forma direta. Não havendo *hard data* que o meça diretamente o processo de interesse, é necessário desenvolver formas de estimar tais parâmetros e estatísticas, sendo este o maior desafio do método proposto.

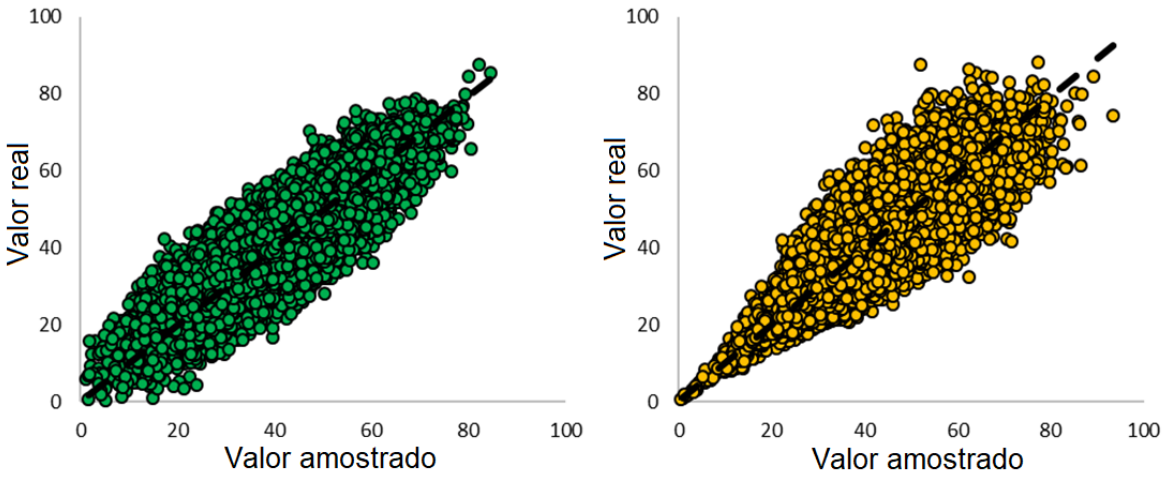
3 SIMULAÇÃO NA PRESENÇA DE ERROS: CASO UNIVARIADO

É impossível definir o desvio exato entre $z_{obs}(\mathbf{u})$ e $z_{real}(\mathbf{u})$, caso contrário, poderíamos simplesmente subtraí-lo e obter o valor real. No entanto, é possível inferir a variância do erro de amostragem $\sigma^2_{erro}(\mathbf{u})$ e possíveis vieses. A seção, a seguir, define um modelo de erros para o caso univariado e, então, são apresentadas alternativas para se estimar esses erros. Portanto, o entendimento do erro associado às observações é um ponto central no desenvolvimento de alternativas geoestatísticas para lidar com observações afetadas por erros.

3.1 MODELANDO O ERRO DE AMOSTRAGEM

A abordagem proposta emprega modelos em que o erro total é gaussiano, sem viés e comumente afetado tanto por erros independentes quanto por aqueles correlacionados com o teor. A validade e generalidade dessas características são discutidas por diversos trabalhos (HOWARTH; THOMPSON, 1976; THOMPSON; HOWARTH, 1973, 1976, 1978). A Figura 4 mostra o comportamento do erro relacionado ao teor (círculos amarelos), em que quanto menor é o teor, menor o erro. Os erros independentes do teor (círculos verdes) mantêm uma distância média constante em torno da linha x = y.

Figura 4 - Comportamento do erro independente e relacionado ao teor



Comportamento das observações afetadas por erros independentes (círculos verdes) e proporcionais ao teor (círculos amarelos). Em preto tracejado a linha x= y.

A Equação 1 define um modelo misto de erro, em que o valor real é afetado por erros absolutos e proporcionais ao teor:

$$z_{obs}(\mathbf{u}) = z_{real}(\mathbf{u}) + \overbrace{\underbrace{A_k X_1(\mathbf{u})}_{\text{Abs. erro Comp.}} + \underbrace{z_{real}(\mathbf{u}_i) C_k X_2(\mathbf{u})}_{\text{Rel. erro Comp.}}}^{\text{Erro total } = \sigma^2_{erro}(\mathbf{u})} \quad (1)$$

O valor real $z_{real}(\mathbf{u})$ é constante em **u**. Portanto, a variância de múltiplas leituras resulta do erro do total, composto por erros vindos do processo de amostragem, preparação e análise

$$\text{Var}\{z_{obs}(\mathbf{u})\} = A_k^2 + z^2_{real}(\mathbf{u}) C^2_k \quad (2)$$

Onde:

$z_{obs}(\mathbf{u})$: observação em **u** da variável k;

$z_{real}(\mathbf{u})$: valor real, o qual é constante em cada posição **u;**

$\sigma^2_{erro}(\mathbf{u})$: variância do erro de amostragem associado a $z_{obs}(\mathbf{u})$;

$X_m(\mathbf{u})$: valores gaussianos $N\{0, 1\}$;

A_k: magnitude do erro Absoluto;

C_k: magnitude do erro Relativo.

A variância global de um conjunto de medições $z_{obs}(\mathbf{u})$ em diferentes posições $\mathbf{u} = 1,\ldots,N$ pode ser encontrada ao substituir o termo $z^2_{real}(\mathbf{u})$ da Equação 2 pela variância global $Z_{real}(\mathbf{u})$:

$$\text{Var}\{z_{obs}(\mathbf{u})\} = A_k^2 + \text{Var}\{Z_{real}(\mathbf{u})\} C^2_k \quad (3)$$

Na seção seguinte, são apresentadas alternativas para estimar o erro amostral. São apresentadas diversas soluções, desde métodos mais simples e que necessita de menos recurso até métodos mais sofisticados, baseados no uso de duplicatas e testes de hipótese. O modelo é estendido ao caso multivariado na Seção 4.1 ESTENDENDO O MODELO DE ERROS PARA O CASO MULTIVARIADO).

3.2 ESTIMANDO O ERRO ASSOCIADO A CADA DADO

Esta subseção apresenta alternativas para inferir os erros: (i) pelo efeito pepita e patamar de variogramas ajustados a diferentes subgrupos; (ii) pela Teoria da Amostragem;

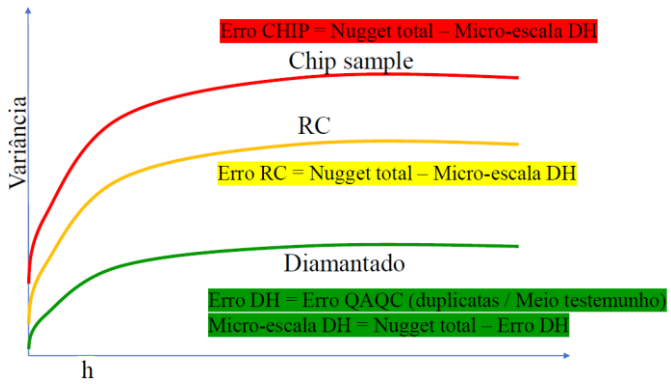
ou (iii) através da variação das leituras entre pares de duplicatas gerados ao longo do processo.

3.2.1 Utilizando o efeito pepita

O efeito pepita é definido como C(**0**) = C_{se} + C_{me}, onde C_{se} é o erro de medição e C_{me} é a variação de microescala (CRESSIE, 1993, p.127-130). A C_{me} depende das características geológicas da amostra e de seu suporte, que pode ser assumida como constante em um banco de dados. A diferença entre o efeito pepita de diferentes grupos é dada pela contribuição do erro C_{se} em suas amostras.

A Figura 5 ilustra a inferência dos erros de amostragem associado a três subgrupos (verde, amarelo e vermelho) através dos seus variogramas. Sendo o C(**0**) dos três modelos obtidos por ajuste de variogramas a esses dados e C_{se} dos furos diamantados (variograma verde) sendo conhecido, o seu C_{me} pode ser obtido. Assumindo a variação de microescala como constante, o erro amostral dos outros dois subgrupos é estimado através de o C_{se} = C(**0**) - C_{me}.

Figura 5 – Inferência dos erros de amostragem de populações através de seus variogramas



Esquema da inferência do erro de amostragem através das componentes do efeito pepita de variogramas ajustados a diversos tipos de dados.

O efeito pepita é uma propriedade isotrópica, portanto, é necessário modelá-lo em apenas uma direção. O variograma médio calculado ao longo dos furos de sondagem (*down-the-hole*) é o mais adequado para capturar as variações a curtas distâncias e, portanto, uma melhor inferência do efeito pepita total. Essa solução é a mais simples, mas têm diversas limitações, entre as principais:

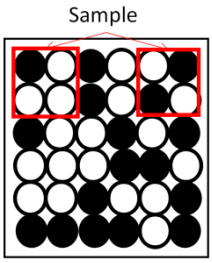
- o método requer que o erro seja independente do teor. No entanto, essa não é a situação mais comum em observações obtidas de a partir de análises químicas feitas em rochas;
- as observações devem ter sido obtidas de suportes amostrais semelhantes. Discrepâncias relevantes no suporte afeta a representatividade do efeito pepita total;
- a definição de subgrupos é geralmente arbitrária, levando a junção de amostras em subgrupos não necessariamente homogêneos.

Na seção seguinte, será discutido como o erro fundamental de amostragem (FSE) pode ser utilizado para estimar o erro associado a cada amostra, em função do seu teor e do protocolo amostral utilizado.

3.2.2 Utilizando a teoria da amostragem

A Teoria da Amostragem (GY, 1982) nos permite quantificar o erro fundamental da amostragem, o qual é devido à heterogeneidade constitucional inerente de materiais e nunca pode ser eliminado. O FSE, chamado de "o menor erro médio residual alcançável" de protocolos amostragem, é inerente à amostra devido a duas características de materiais de minério fragmentados: a heterogeneidade composicional e a heterogeneidade da distribuição espacial dos elementos (Figura 6).

Figura 6 – Exemplo da heterogeneidade constitucional de um lote



Esquema de como a heterogeneidade constitucional afeta o erro de amostragem. Os quadrados representam a área de delimitação para coletas de amostras do mesmo lote, mas que levam a dois resultados diferentes devido à heterogeneidade do lote.

A heterogeneidade é expressa pelo fator de forma (*f*), pelo fator granulométrico (*g*), pelo fator (*c*) de composição mineralógica e pelo fator de liberação (Equação 4):

$$\sigma^2_{fse} = \left(\frac{1}{M_S} - \frac{1}{M_L}\right) clfgd^3 \qquad (4)$$

onde σ^2_{fse} é a variância do erro, M_S e M_L são as massas da amostra e do lote, respectivamente. Discussões sobre como definir os fatores c, l, f, g e d discutidos em detalhe em Gy (1982).

A equação 4 é consistente com o modelo de erros proposto (Equação 1), uma vez em que os fatores *g* e *c* são proporcionais ao teor e d^3 é constante. Assumir que o FSE é o erro total associado a uma dada observação só é valido quando todas as outras fontes de erros podem ser assumidas como nulas, o que, geralmente, não é o caso. Mudanças, adaptações e alterações do processo amostral que não são documentadas, assim como mudanças despercebidas das características materiais, adicionam novas componentes de variação que modificam o erro esperado associado ao processo de amostragem e análise. Nestas condições, pode ser mais apropriado o uso de dados experimentais de controle de qualidade para estimar o comportamento do modelo utilizando dados de duplicatas.

3.2.3 A variância de krigagem como erro amostral

Deraisme; Strydom (2009) definem o erro amostral de dados de baixa qualidade através da variância de estimativa de CoK em suas posições. O método traz limitações inerentes CoK e na decisão de assumir a variância de estimativa como erro de medição. Na prática, faz com que todos os dados sejam agrupados em até 3 subgrupos devido a quantidade de variogramas diretos e cruzados a se modelar e as restrições impostas à definição do modelo linear de corregionalização. Nesse caso, quão maior a quantidade de subgrupos agrupados para a modelagem de um único modelo variográfico, maiores as perdas de acurácia do modelo.

O segundo problema é que existem propriedades da variância de krigagem que são incorretas de se atribuir como uma propriedade ao erro amostral, como:

- a variância de krigagem não depende dos valores das amostras, e sim do variograma e da configuração espacial dos dados. Portanto, é assumido que o erro é independente do teor;
- a variância de krigagem cresce quando a distância entre a observação e o ponto a ser estimado aumenta. Isso significa que quão mais longe uma amostra está de outras, maior o erro amostral atribuído a ela.

A primeira propriedade limita as aplicações a apenas algumas situações específicas, já que é comum haver correlação entre erro amostral e teor. A segunda propriedade faz com que o sistema de krigagem puna a observação em função da sua posição duas vezes. O quão mais distante uma dada observação esteja de outras, menor o peso que o sistema de krigagem atribui a elas. E ao adicionar a variância do erro nas posições de observações (**h** = 0) em que tal variância deveria ser zero, essa adição diminui ainda mais o peso atribuída a essa observação, e, portanto, a redução do peso atribuído a essa amostra é desproporcional.

3.2.4 Definindo subgrupos de erro através de amostras duplicatas

Esta seção propõe a abordagem para definir o erro associado a cada amostra publicada no artigo: "*Using QC data to estimate the individual precision of original samples: a duplicates-based approach*" de Silva; Costa (2018).

O erro individual associado a cada amostra é estimado pela equação ajustada ao subgrupo ao qual pertence. Subgrupos são aqui definidos como conjuntos de amostras produzidas pelas mesmas condições e afetadas pelas mesmas fontes de erros. As variações dentro de cada subgrupo devem resultar, principalmente, de efeitos aleatórios, enquanto as variações entre subgrupos devem ser causadas por diferenças no processo amostral ou nas características do material. A definição de subgrupos e suas equações de erro são baseadas no desvio entre pares de duplicatas em sucessivas etapas de análise gráfica e testes de significância estatística. Dados como tipo de rocha ou os protocolos de amostragem utilizados podem apoiar na definição de subgrupos. As principais vantagens da abordagem proposta são:

- é capaz de avaliar o impacto das informações disponíveis sobre o processo, por exemplo, de qual método de sondagem origina-se cada dado. É também possível definir subgrupos sem ser necessário conhecer as causas que afetam a formação do grupo;
- uso de sucessivos testes de hipóteses para avaliar cada possível subgrupo torna o método menos subjetivo;
- pode ser facilmente implementado em qualquer software de planilha.

A metodologia proposta é composta pelas seguintes etapas:

i. Testando critérios que definam subgrupos: a capacidade de informações como tipo de rocha, suporte de amostragem ou protocolos utilizados para definir subgrupos são analisadas. Para cada um desses critérios, uma equação linear é ajustada aos seus valores de duplicatas, relacionando o teor e o erro de medição. O erro estimado pela equação é comparado aos erros experimentais de cada par de duplicatas, sendo esse desvio entre o valor estimado e o valor experimental, chamado de resíduo. O subgrupo testado é considerado como representativo quando $Resíduo_{individual} < Resíduo_{geral}$ é estatisticamente significativo, indicando que os ganhos do ajuste da equação, a apenas aquele conjunto de dados, têm erros de regressão significantemente menores do que um ajuste mais geral, sem essa subdivisão testada.

ii. Testando a ocorrência temporal de subgrupos: a ordem temporal de preparação ou análise nos permite detectar flutuações do erro que ocorreram ao longo do tempo (MONTGOMERY, 2009, p.193). Porntanto, na segunda etapa, cada subgrupo estatisticamente significativo tem o desvio de seus pares de duplicatas traçado em um gráfico cumulativo de soma (CUSUM, do inglês *Cumulative Sum*) ordenados de forma temporal.

O mesmo teste de hipótese utilizado na etapa (i) é aplicado aos (*i, j, k...n*) subgrupos individualizados. O erro modelado para cada subgrupo é testado dois a dois para avaliar seus resíduos. Por exemplo, se o teste $Resíduo_i < Resíduo_{i+j}$ falhar, o par é fundido i + j e, em seguida, é testado contra um segundo conjunto, no caso se testaria $Resíduo_{i+j} < Resíduo_{i+j+k}$.

iii. Estimando o erro individual das amostras: a incerteza de medição para as amostras originais é estimada em função dos seus teores individuais através do modelo de erro ajustado às duplicatas de cada subgrupo estatisticamente significativo.

Embora as etapas da metodologia possam ser aplicadas a qualquer tipo de duplicatas, o uso de amostras mais próximas da etapa de coleta de campo tem uma maior chance de medir mudanças de processo que definam subgrupos. Por exemplo, enquanto um par de amostras geradas na subamostragem de polpa só será afetado pelas etapas subsequentes, as réplicas de campo coletadas em conjunto com a amostra original são afetadas por todas

as possíveis fontes de erros de precisão ao longo de todo o processo. As etapas são discutidas em maior detalhe a seguir.

3.2.4.1 Definição de candidatos do subgrupo utilizando informações prévias

O primeiro passo para delimitar subgrupos é avaliar quais das informações disponíveis podem indicar subgrupos. São tipos de informações relevantes os litotipos, método de análise química, métodos e massas das etapas primárias e sub amostragem etc. As principais fontes de definição de subgrupos resultam de três causas principais (GY 1982; ABZALOV 2011, p. 611-612):

- erros de agrupamento-segregação e erros durante o procedimento de extração e preparação da amostra -essas fontes de erros são controladas pela forma e tamanho das partículas, o tamanho em que os componentes críticos são liberados e as diferenças de mineralogia e densidade entre a ganga e os minerais minério;
- erros que dependem de quão rigoroso o protocolo de amostragem foi desenvolvido, implementado e seguido – são causados pela extração incorreta de amostras, procedimentos de preparação abaixo do ideal, contaminação ou erros humanos;
- erros analíticos e instrumentais que ocorrem durante as operações analíticas, incluindo o ensaio, análise de umidade, pesagem de alíquotas, análise de densidade, erros de precisão e viés causado pelo baixo desempenho de instrumentos analíticos.

Após definir possíveis critérios que delimitem subgrupos, é necessário medir quais geram modelos que sejam estatisticamente significantes. Esse teste é que torna o método não arbitrário. A maioria dos métodos para definição de *hard data* e subgrupos de *soft data* para nessa etapa, em que um dado critério é empregado sem testes que garantam a relevância.

3.2.4.2 Modelar o comportamento de erro (método de Thompson-Howarth)

Quando há um grande faixa de valores nos teores de um conjunto de amostras, os erros absolutos e relativos podem variar significativamente ao longo do intervalo, e, assim, o desvio padrão não é capaz de descrever adequadamente a precisão das observações (THOMPSON; HOWARTH, 1976). A forma mais adequada de análise seria usando uma equação que relacione o erro ao teor (Figura 7). O método se baseia nas premissas de que

os dados não contem viés e que o erro de amostragem tem uma distribuição Gaussiana com a variância função do teor.

A relação $\sigma^2_{erro}(.) = A_k + z_{real}(.)C_k$ (Equação 2) relaciona o desvio-padrão entre duplicatas para um dado teor com a soma da constante A_K que mede o desvio padrão quando o teor é zero e o produto entre a inclinação C_K e o teor $z_{real}(.)$ médio do par de duplicatas. A abordagem proposta por Thompson e Howarth para definir as constantes da Equação 2 requer pelo menos 50 pares de duplicatas e é composta pelos seguintes passos:

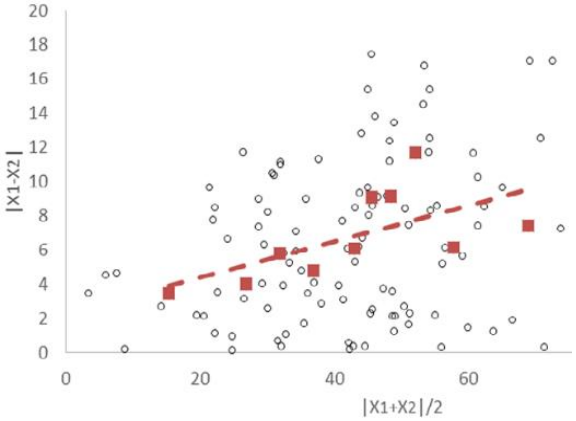
i. as diferenças relativas entre pares de duplicatas são calculadas por $2|X_1 - X_2|/(X_1 + X_2)$;

ii. os pares de dados são ordenados ascendentes por seus teores médios e subdivididos em grupos a cada 11 resultados consecutivos;

iii. calcula-se o teor médio dos 11 pares de amostras de cada grupo de dados e a mediana das suas diferenças absolutas;

iv. A mediana do erro é então traçada em função do teor médio do grupo. Uma regressão linear é ajustada aos pontos experimentais.

Figura 7 – Relação erro X teor



Regressão da mediana das diferenças absolutas (y) em função do teor médio (x). Círculos são resultados de cada par individual, quadrados vermelhos é a mediana de grupos de 11 pares consecutivos. A linha vermelha é a regressão linear.

Várias abordagens foram desenvolvidas para estimar da relação entre o erro e o teor quando as suposições de Thompson-Howarth não são aplicáveis, como quando os erros são não-gaussianos (STANLEY; LAWIE, 2008), variáveis geoquímicas exibindo um comportamento com efeito pepita elevado (STANLEY, 2006), ou quando é necessário

separar as variações de amostragem e medições para identificar a forma de sua relação contra o teor (FRANÇOIS-BONGARCON, 1998). A seleção do modelo precisa ser rigorosamente testada para banco de dados e subgrupos. O modelo entre erro e teor utilizado pode ser facilmente substituído por qualquer outro sem afetar as outras etapas do fluxo de trabalho proposto.

3.2.4.3 Verificando a significância estatística dos subgrupos

Para testar se o comportamento da variância do erro de uma dada amostra, em função do seu teor z_{real}, o erro $\sigma^2_{erro}(.; z_{real})$ é modelado para um determinado subgrupo é estatisticamente significativo, são verificados se os resíduos da equação ajustada a esses subgrupos têm resíduos significativamente menores do que aquele obtido por uma equação ajustado aos dados sem essa subdivisão. A significância estatística é medida pelo teste-t emparelhado (MONTGOMERY; RUNGER, 2014, p. 400-401). Esse é um caso especial do "teste t de duas amostras", que consiste em analisar as diferenças dos resíduos para cada par de duplicatas, de forma a definir os parâmetros da Equação 2, que relaciona o erro ao teor. Seja (X_{11},X_{21}), (X_{12},X_{22}), ... (X_{1n}, X_{2n}) um conjunto de n duplicatas, então a diferença para cada par é definida por

$$D_J = X_{1j} - X_{2j}, \text{ onde } j = 1, 2,..., n \quad (5)$$

onde D_j é assumido como normalmente distribuído com média $\mu_D = E(X_1 - X_2) = E(X_1) - E(X_2) = \mu_1-\mu_2$. Assim, o teste da significância da diferença para μ_1 e μ_2 (Equação 6) é realizado por um teste-t de uma amostra em μ_D, ou seja, se testa se H_0: $\mu_{geral} - \mu_{individual} = 0$ contra H_1: $\mu_{geral} - \mu_{individual} > 0$, o que é equivalente ao teste:

$$\begin{aligned} H_0&: \mu_d = 0 \\ H_1&: \mu_d > 0 \end{aligned} \quad (6)$$

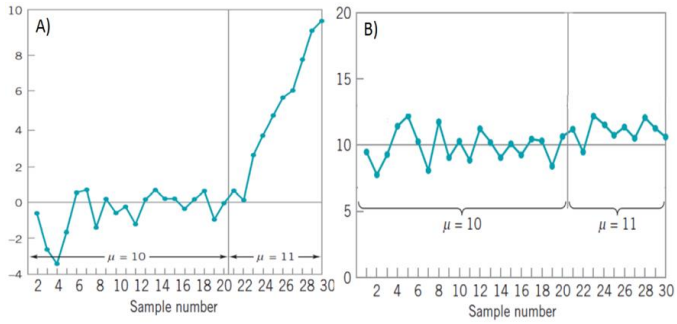
Embora o teste-t emparelhado seja formalmente baseado na suposição de gaussianidade, bons resultados com dados não gaussianos são obtidos se o número de amostras disponíveis for suficientemente grande. A relação entre robustez e gaussianidade e número de amostras depende de quão não normais são os dados. Um conjunto de 20 pares de amostras é geralmente o bastante (MINITAB, 2017). Considerando que o método proposto requer pelo menos 50 amostras para definir a linha de regressão, o teste-t emparelhado é apropriado.

3.2.4.4 Segunda etapa da definição de subgrupos: O gráfico de soma cumulativa

É relevante a possibilidade de o método proposto permitir definir subgrupos sem conhecer com antecedência as causas ou fontes de variação. Do ponto de vista prático, bancos de dados reais são comumente compostos por subgrupos resultantes de diferentes variações de processos, sendo elas conhecidas ou não.

Essa etapa da definição de subgrupos será baseada no uso de gráficos de control, em que os desvios de um valor-alvo são acumulados em uma soma. A maior sensibilidade do gráfico CUSUM (Figura 8a) deve-se à incorporação de todas as informações em uma sequência de valores amostrais, tornando esse gráfico capaz de detectar mudanças de cerca de 1.5 desvio-padrão. Essa solução é mais sensível do que métodos como o gráfico de Shewhart (Figura 8b) que detectam apenas mudanças superiores a dois desvios-padrão. A ordem de preparação ou análise permite detectar flutuações que ocorrem ao longo do tempo (MONTGOMERY, 2009).

Figura 8 – Comparativo da carta de controle de Shewart e da CUSUM



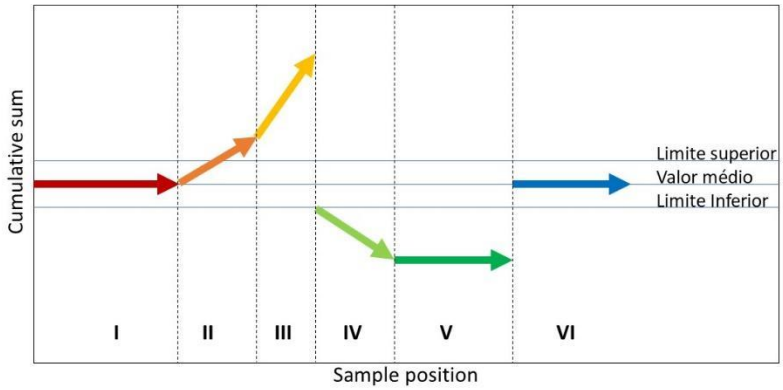
a) CUSUM e b) Carta de controle de Shewhart para os mesmos dados - a linha vertical indica onde a média do processo muda de μ = 10 ao μ = 11. Adaptado de "Controle estatístico de qualidade", por Montgomery (2009).

O funcionamento do CUSUM é baseado na acumulação dos desvios que estão acima do alvo, chamado de C+, assim como acumulando desvios que estão abaixo da meta através de C-, ambas partindo do zero. Enquanto a média do processo estiver ajustada ao valor alvo, a função oscilará aleatoriamente em torno de zero, pois os desvios positivos serão anulados pelos negativos. Se a média do processo aumenta (ou diminui), a linha aumentará (ou diminuirá) indefinidamente. Os valores C+ e C-, chamadas respectivamente como uma face superior e inferior, são computadas da seguinte forma:

$$C^{+}_{i} = \max[0, X_i - (\mu_0 + R) + C^{+}_{i-1}]$$
$$C^{-}_{i} = \max[0, (\mu_0 - R) - X_e + C^{-}_{i-1}] \quad (7)$$

Onde X_i é a amostra no tempo i, μ_0 é a média global das amostras e R é o valor alvo. É importante entender o verdadeiro significado do comportamento traçado pelo CUSUM para definir candidatos a subgrupos. Uma inclinação constante (para baixo ou para cima) indica que o erro do subgrupo também é constante, e que mudanças na ordem de grandeza do erro causam mudanças na inclinação da reta (ou no seu sentido) (Figura 9).

Figura 9 - Indicadores de mudanças do erro médio



Indicadores de mudanças na magnitude do erro em função de variações na inclinação do desvio acumulado do CUSUM. Cores diferentes indicam os candidatos a subgrupos de I a VI.

Outro ponto é que há uma grande probabilidade de alarmes falsos dentro das amostras iniciais porque o erro cumulativo de alguns valores é menos robusto para variações aleatórias.

3.2.4.5 estimando a precisão de medição de cada amostra original

Na última etapa, a relação entre a variância do erro amostral e o teor de cada observação $\sigma^2_{erro}(. \; ; z_{real})$ é definida para cada subgrupo estatisticamente significativo e usada para estimar o erro de cada amostra original. Além disso, o erro da regressão linear de cada subgrupo pode ser usado para definir os intervalos de confiança para esses modelos, bem como para medir o intervalo de confiança de cada valor estimado (MONTGOMERY; RUNGER, 2014; MONTGOMERY, 2009).

3.3 EXEMPLO 1 - ESTIMANDO OS ERROS DOS DADOS

Para exemplificar a metodologia proposta por Silva; Costa (2018), usamos um gerador pseudoaleatório para simular um conjunto de dados típico de duplicatas. Por uma questão

de simplicidade, presume-se que todos os dados são medições feitas com amostras do mesmo tipo de rocha e não há informações disponíveis sobre mudanças que ocorreram no processo de preparação e análise ao longo do tempo.

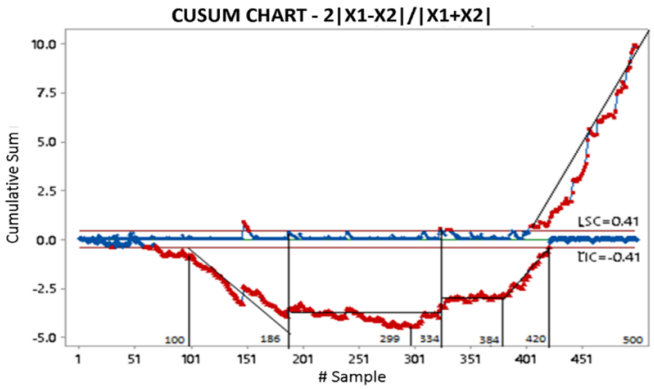
Foram simulados pares de amostras imitando duplicatas. O valor $z_{real}(\mathbf{u})$ de cada amostra original foi simulado pela tiragem de 500 valores de uma distribuição $N\{40, 15\}$, assumindo-os como sendo valores do fenômeno real, Z_{real}. Cada amostra de um par de duplicatas é, então, gerada pela adição de um novo valor de erros tirado de forma independente de $N\{z_{real}, \sigma^2_{erro}\}$.

Tabela 1 - Intervalos e parâmetros dos subgrupos sintéticos

Subgrupo	Intervalo	A_k	C_k
A	1-100	2	0.15
B	101-190	2	0.08
C	191-300	3	0.16
D	301-420	4	0.20
E	421-500	1	0.40

A primeira definição de candidatos a subgrupos é realizada ao traçar, em um gráfico CUSUM, o erro relativo $2|X_1 - X_2|/(X_1 + X_2)$ de pares correspondentes de duplicatas organizados temporalmente. Com base na tendência e inclinação dos valores fora de controle (Figura 10, pontos vermelhos), foram definidos sete subgrupos para terem suas significâncias testadas (1, 2, 3... 7).

Figura 10 – CUSUM ajustado aos dados simulados com erro



CUSUM do desvio medido por valor $2|\ X_1\text{-}X_2|/\ (X_1+X_2)$ nos dados sintéticos. As linhas verticais delimitam possíveis subgrupos definidos pela análise gráfica. As linhas horizontais indicam os intervalos delimitados. O limite superior (LSC) e inferior (LIC) referem-se aos intervalos de controle. Pontos vermelhos indicam valores fora dos intervalos de controle; pontos azuis indicam valores dentro dos limites.

Cada candidato a subgrupo foi testado para avaliar se o resíduo da equação ajustada aos seus dados $\mu_{individual}$ é estatisticamente menor do que o resíduo de uma equação mais geral μ_{geral}. O resíduo corresponde ao desvio entre o erro estimado para pares de duplicatas através da equação modelada e o erro experimental medido para esse mesmo par.

Por exemplo, a primeira linha da Tabela 2 tese a hipótese de que $\mu_{eq\,(1)} < \mu_{eq\,(1+2)}$. O valor *p* resultante para o subgrupo 1 foi de 88.6%, o que significa que não há evidências, para um nível de significância de 90%, que a definição de uma linha de regressão apenas para o subgrupo 1 reduziria os resíduos da regressão. Em oposição, a probabilidade de que $\mu_{eq(2)}<\mu_{eq\,(1+2)}$ é de 99.8%, o que sugere que a equação deva ser modelada separadamente, pois há evidencia que levaria a estimativas de melhor qualidade para o subgrupo 2.

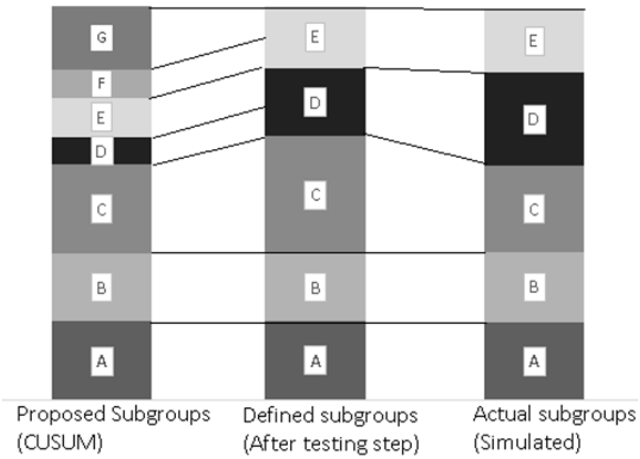
Tabela 2 – Série de testes de hipótese (teste-t emparelhado)

Subgrupo	Intervalo	Teste 1	p-value 1	Teste 2	p-value 2	Subgrupo definido
1	1-100	$\mu_{eq\,(1+2)}-\mu_{eq\,(1)}>0$	88.6	$\mu_{eq\,(1+2)}-\mu_{eq\,(2)}>0$	**99.8**	A (1-100)
2	101-186	$\mu_{eq\,(2+3)}-\mu_{eq\,(2)}>0$	**99.8**	$\mu_{eq\,(2+3)}-\mu_{eq\,(3)}>0$	81	B (101-186)
3	187-299	$\mu_{eq\,(3+4)}-\mu_{eq\,(3)}>0$	60.4	$\mu_{eq\,(3+4)}-\mu_{eq\,(4)}>0$	54.5	C (187-334)
4	300-334	$\mu_{eq\,(3+4+5)}-\mu_{eq\,(3+4)}>0$	72.3	$\mu_{eq\,(3+4+5)}-\mu_{eq\,(5)}>0$	**99.2**	
5	335-384	$\mu_{eq\,(5+6)}-\mu_{eq\,(5)}>0$	80.3	$\mu_{eq\,(5+6)}-\mu_{eq\,(6)}>0$	70.5	D (335-420)
6	385-420	$\mu_{eq\,(5+6+7)}-\mu_{eq\,(5+6)}>0$	84.9	$\mu_{eq\,(5+6+7)}-\mu_{eq\,(7)}>0$	**99.2**	
7	421-500	-	-			E (421-500)

Avaliou-se se os resíduos de regressão usando amostras de um único subgrupo são estatisticamente inferiores a uma regressão geral que mescle subgrupos sucessivos ($\mu_{general} - \mu_{individual} >0$) com um nível de significado de 90%. *p-value* negritados indicam quais subgrupos individuais levam a melhores estimativas de erro.

Após a definição gráfica dos candidatos do subgrupo e da série de testes de hipótese (Tabela 2), os subgrupos 3-4 e 5-6 foram fundidos devido à falta de significância de suas diferenças. O resultado ficou muito próximo daquele real que foi originalmente simulado (Figura 11).

Figura 11 – Comparação entre os subgrupos reais e estimados



Comparação entre os subgrupos inicialmente definidos pelo método proposto; resultados após o teste de hipótese e os subgrupos reais.

O desvio relativo médio do erro estimado pelo método proposto foi de 11.40%. Os valores de outras três abordagens são:

- Definição de uma única equação a todos os dados com erro de 16.3%;
- Definindo um único erro relativo médio para todas as amostras com erro de 19.6%;
- Erro relativo definido por média móvel de 10 valores (em ordem de tempo) tem um erro médio de estimativa de 17.10%.

Comparando o método proposto com o segundo com melhor desempenho, que foi o uso de Thompson-Howarth sem subdivisão de subgrupos, o primeiro apresentou uma estimativa de quase 30% menos erro.

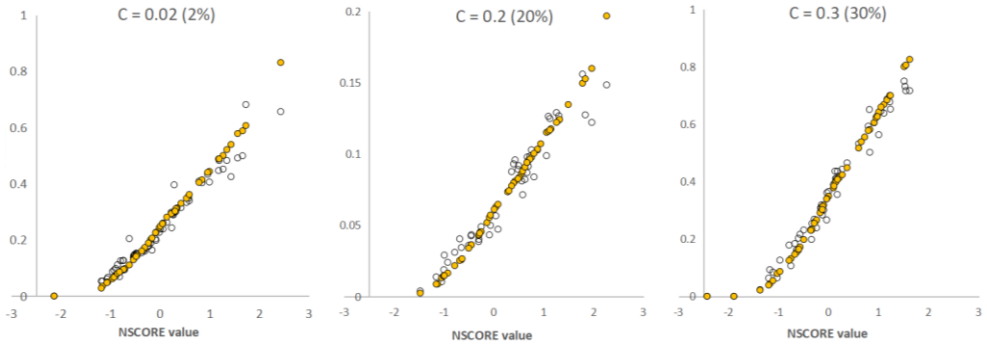
3.4 TRANSFORMANDO OS ERROS PARA VALORES ESTANDARIZADOS

O uso de métodos geoestatísticos de simulação baseados na hipótese multiGaussiana, torna necessário transformar os valores iniciais para gaussianos *N*{0,1}. Nessa situação, erros dependentes do teor devem ser transformados em valores independentes.

É apresentada uma solução analítica e uma numérica para converter os erros amostrais, sejam eles independentes ou proporcionais ao teor, de suas unidades originais para valores Gaussianos. A Figura 12 compara as abordagens através de dados sintéticos, os valores dados com diferentes erros (Equação 1 usando A = 0 e C_k = 0.02, 0.2 ou 0.3) são transformados para Gaussianos. Na Figura 12, os valores obtidos analiticamente (círculos

laranja) mostram uma boa aderência com os obtidos numericamente (círculos pretos vazios).

Figura 12 – Erro em Normal-score estimado por método analítico e computacional



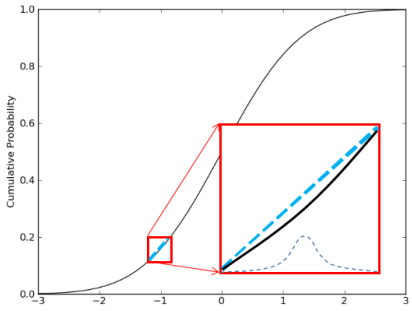
Valores da variável de erro proporcional inferidas pela solução computacional (círculos pretos vazis) e analiticamente (círculos laranja). A magnitude dos erros gaussianos é calculada para valores perturbados por erros sorteados em unidades originais usando C_k de 0.1, 0.2 e 0.3.

As duas soluções a seguir lidam com erros dependentes ou independentes do teor. Para tal, todos os componentes (independentes ou proporcionais ao teor) são respectivamente colapsados em uma componente independente A_K e um componente proporcional C_K. A proporção de cada componente na variância total é a mesma em unidades originais e gaussianas. Portanto, a variância total da componente transformada para unidades gaussianas O_k e P_k é redistribuída proporcionalmente aos componentes A_K, B, C_K e D nas suas contribuições em unidades originais à variância total.

3.4.1 Transformando os erros para unidades padronizadas - abordagem computacional

Uma solução computacional pode ser usada para transformar os erros para unidades gaussianas. A abordagem baseia-se na suposição de que o intervalo da CDF que corresponde ao valor e seu espaço de incerteza pode ser assumido como linear. O método discretiza a CDF da distribuição gaussiana de referência $N\{0,1\}$ em seguimentos lineares com o comprimento da distribuição dos possíveis valores que cada observação poderia assumir (Figura 13).

Figura 13 – Distribuição cumulativa discretizada em retas



Distribuição cumulativa usada para a transformação gaussiana. Detalhe mostrando discretização linear (linha azul tracejada) do intervalo de possíveis valores de uma determinada observação.

A transformação dos erros é realizada individualmente para cada valor original $z_{obs}(\mathbf{u})$ da seguinte forma:

i. Para cada observação em **u**, realizações (L) $z^{L}_{obs}(\mathbf{u})$ são simuladas. Sendo o valor real $z_{real}(\mathbf{u})$ desconhecido, todas as realizações são sorteadas de $N\{z_{obs}(\mathbf{u}), z_{obs}(\mathbf{u})*\sigma_{erro}(\mathbf{u})\}$.

ii. Todos os valores $z^{L}_{obs}(\mathbf{u})$ são transformados para gaussianos $y^{L}_{obs}(\mathbf{u})$ usando a transformada $F^{-r}_{Y}(F_Z(z_{obs}))$, que é a mesma distribuição de referência dos valores e do seu erro nas unidades originais.

iii. Em cada posição **u**, calcular o desvio padrão do conjunto de valores transformados.

A distribuição transformada $z^{L}_{obs}(\mathbf{u})$ tem a mesma média e desvio padrão da distribuição $N\{y_{obs}(\mathbf{u}), \sigma_{y(erro)}(\mathbf{u})\}$. Esse método não requer o conhecimento das componentes do modelo de erro, baseando-se no valor de erro associado à observação. Portanto, ele consegue lidar com qualquer tipo de modelo de erros, mesmo os mais complexos.

3.4.2 Transformando os erros para unidades padronizadas - abordagem analítica

O componente do erro proporcional $C_kX_m(\mathbf{u})$ e o independente $A_kX_m(\mathbf{u})$ podem ser transformados analiticamente para unidades gaussianas. Quando a componente independente A_k for nula, a Equação 1 pode ser simplificada para:

$$z_{obs}(\mathbf{u}) = z_{real}(\mathbf{u}) + z_{real}(\mathbf{u})C_kX_2(\mathbf{u}) \quad (8)$$

A componente proporcional é transformada em um novo modelo em que $y_{obs}(\mathbf{u})$ é o valor observado transformado para gaussiano e as observações são compostas pelo valor gaussiano $y_{real}(\mathbf{u})$ mais um erro absoluto O_k.

$$y_{obs}(\mathbf{u}) = y_{real}(\mathbf{u}) + O_k X_2(\mathbf{u}) \quad (9)$$

O valor tirado aleatoriamente de $X_2(\mathbf{u})$ é o mesmo nas unidades originais e transformadas, visto que a transformação não é uma nova tiragem aleatória. Para definir a equivalência entre eles, resolvemos a Equação 8 e Equação 9 para $X_2(\mathbf{u})$

$$X_2(\mathbf{u}) = \frac{z_{obs}(\mathbf{u}) - z_{real}(\mathbf{u})}{z_{real}(\boldsymbol{u}_i) C_k}$$

$$X_2(\mathbf{u}) = \frac{y_{obs}(\mathbf{u}) - y_{real}(\mathbf{u})}{O_k}$$

Reorganizando a igualdade acima, aplicamos Var{} em ambos os lados e substituímos $z_{real}(\mathbf{u})$, que é desconhecido, pela observação disponível $z_{obs}(\mathbf{u})$ já que $E\{Z_{real}(\mathbf{u})\} = z_{obs}(\mathbf{u})$. Isso leva a:

$$O_k(\boldsymbol{u}_i) = \sigma\{y_{obs}(\boldsymbol{u}_i)\} \frac{C_k z_{obs}(\boldsymbol{u}_i)}{\sigma\{z_{obs}(\boldsymbol{u}_i)\}} \quad (10)$$

Em palavras, a Equação 10 mostra que a proporção entre a componente total de erro e a variância global é mantida na transformação das unidades originais para gaussianas. A definição da magnitude da componente em unidades gaussianas quando a componente proporcional C_k é nula é semelhante:

$$P_k(\mathbf{u}) = A_k \frac{\sigma\{y_{obs}(\mathbf{u})\}}{\sigma\{z_{obs}(\mathbf{u})\}} \quad (11)$$

Devido a substituição de $z_{real}(\mathbf{u})$ por $z_{obs}(\mathbf{u})$, quanto maior for o desvio entre esses valores, maior será o erro causado por essa solução. A inspeção visual do erro gaussiano em função da observação transformada (Figura 12, círculos laranja) mostra o impacto dessa solução. Caso seja necessário, o ajuste de uma equação a esses valores suaviza essas variações aleatórias.

3.5 SIMULANDO O PROCESSO REAL A PARTIR DE OBSERVAÇÕES: CASO UNIVARIADO

O fluxo de trabalho convencional tem suas realizações condicionadas a reproduzir o variograma e histograma ajustado às observações. Como resultado, as realizações superestimam a variabilidade tanto local quanto global do teor e subestimam a incerteza do modelo nos locais amostrados ao ignorar a variância do erro que é associada aos dados. Tal fluxo é dado pelos seguintes passos: (i) transformação das observações para unidades gaussianas, (ii) ajuste do variograma para os valores transformados, (iii) simulação das realizações, (iv) retro transformação das simulações para suas unidades originais.

Como alternativa, a abordagem proposta infere o histograma e o variograma do fenômeno real a ser simulado. As observações são substituídas por realizações de *hard data*. Os bancos de dados sem erros e os parâmetros inferidos condicionam as realizações de todo o modelo.

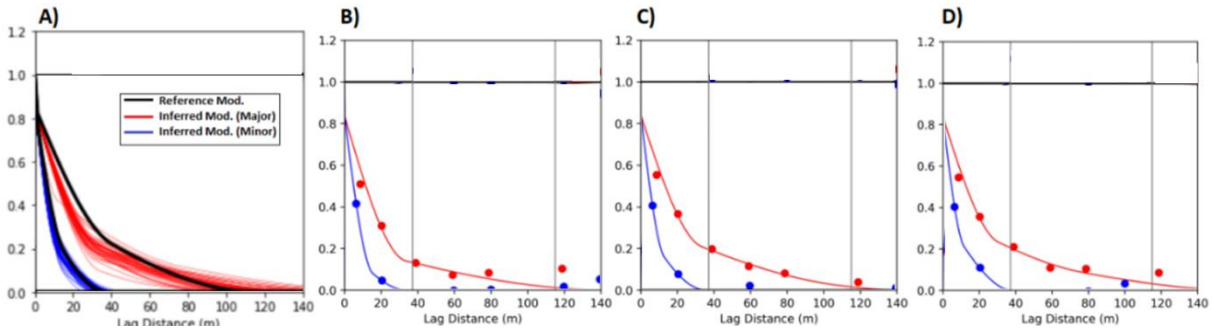
3.5.1 Metodologia

O variograma do fenômeno real é inferido através dos modelos ajustados às observações e os seus erros. Em aplicações reais, o comportamento $\varepsilon(\mathbf{u})$ pode ser estimado por um dos diferentes métodos apresentados (Seção 3.1 MODELANDO O ERRO DE AMOSTRAGEM). Assumindo que estimativas dos valores de $\varepsilon(\mathbf{u})$ estejam disponíveis, o método pode ser divido em duas etapas principais:

i. os modelos do histograma e variograma do fenômeno real são inferidos a partir dos modelos ajustados às observações e as estimativas dos erros individuais $\sigma^2_{erro}(\mathbf{u})$;

ii. parâmetros inferidos e erros de amostragem são usados para simular possíveis *hard data* equiprováveis ao fenômeno real. As observações são substituídas pelas realizações de possíveis valores livres de erros $z^L_{real}(\mathbf{u})$. A CoK estima a distribuição cumulativa local em cada posição utilizando observações colocalizados, não colocalizados e os *hard data* previamente simulados. Os conjuntos de *hard data* simulados honram os parâmetros inferidos e são representativos do fenômeno real (Figura 14).

Figura 14 – Covariograma de referência e valores simulados



A) modelo do covariograma real (linha preta) e os modelos resultantes de 30 realizações (linhas azuis e vermelhas). As Figuras B), C) e D) apresentam o covariograma modelado experimental de três realizações diferentes. Linhas cinzas verticais mostram o alcance do modelo de referência).

O espaço de incerteza entre realizações em um nó amostrado **u** é controlado, principalmente, pelo erro de amostragem $\sigma^2_{erro}(\mathbf{u})$ associado à observação colocalizada $z_{obs}(\mathbf{u})$. Os valores simulados são assumidos como realizações do fenômeno real nas posições amostradas. As realizações de $Z_{real}(\mathbf{.})$ são condicionadas a honrar cada valor simulado e os parâmetros inferidos. Como todos os valores utilizados nessa etapa são assumidos como *hard data*, pode-se empregar métodos de simulação estocástica mais simples, tais como SGS, com a estimativa das probabilidades locais através de SK. Para cada realização simulada, um novo banco de dados deve ser simulado e considerado.

Na presença de erro de amostragem, a distribuição e a correlação espaciais ajustadas às observações podem não ser representativas do comportamento do fenômeno real. A variância medida na distribuição e o variograma (ou outra medida de correlação espacial) combinam o comportamento do fenômeno de interesse com o do erro associado ao processo de medição. A seguir, métodos para estimar o covariograma e a distribuição do fenômeno real serão apresentados.

3.5.1.1 Estimando o covariograma do fenômeno real - erro independente do teor

Erros independentes do teor não alteram o covariograma, pois mantêm constante diferença entre efeito pepita C(**0**) e patamar. Erros independentes do teor adicionam a mesma variação de erro $\sigma^2_{erro}(\mathbf{u})$ ao efeito pepita e na contribuição de cada estrutura e ao patamar. Com base nas relações $C_{obs}(\mathbf{h}) = \sigma^2_{obs} - \gamma_{obs}(\mathbf{h})$, $\sigma^2_{obs} = \sigma^2_{erro} + \sigma^2_{real}$ e $\gamma_{obs}(\mathbf{h}) = \gamma_{real}(\mathbf{h}) + \sigma^2_{erro}$, temos:

Prova 1:

$$C_{obs}(\mathbf{h}) = \sigma^2_{obs} - \gamma_{obs}(\mathbf{h})$$
$$= C_{obs}(\mathbf{h}) = \sigma^2_{erro} + \sigma^2_{real} - (\gamma_{real}(\mathbf{h}) + \sigma^2_{erro})$$
$$= C_{obs}(\mathbf{h}) = \sigma^2_{real} - \gamma_{real}(\mathbf{h})$$
$$= C_{real}(\mathbf{h})$$

A mesma equivalência $Cov\{z_{obs}(\mathbf{u}), z_{obs}(\mathbf{u})\} = Cov\{z_{real}(\mathbf{u}), z_{real}(\mathbf{u})\}$ também pode ser provada de forma mais geral através das propriedades de bilinearidade da covariância, a qual é uma importante generalização da expansão quadrática. Simplificando a notação onde Cov{a,b} = [a,b], essa propriedade é dada por:

$$Cov\{X + Y, X - Y\} = [X + Y, X - Y] = [X, X] - [X, Y] + [Y, X] - [Y, Y] \quad (12)$$

Utilizando essa propriedade-chave da covariância, podemos desenvolver a prova 2:

Prova 2:

$Cov\{z_{real}(\mathbf{u}), z_{obs}(\mathbf{u})\}$
$= [z_{real}(\mathbf{u}), z_{real}(\mathbf{u}) + \sigma^2 Z_{(erro)}(\mathbf{u})]$
$= E\{Z_{real}(\mathbf{u})[Z_{real}(\mathbf{u}) + \sigma^2 Z_{(erro)}(\mathbf{u})]\} - E\{Z_{real}(\mathbf{u}) + \sigma^2 Z_{(erro)}(\mathbf{u})\}E\{Z_{real}(\mathbf{u}) + \sigma^2 Z_{(erro)}(\mathbf{u})\}$
$= E\{Z_{real}(\mathbf{u})\sigma^2_{erro}\{\mathbf{u}\} + Z_{real}(\mathbf{u})Z_{real}(\mathbf{u})\} - E\{Z_{real}(\mathbf{u}) + \sigma^2 Z_{(erro)}(\mathbf{u})\}E\{Z_{real}(\mathbf{u}) + \sigma^2 Z_{(erro)}(\mathbf{u})\}$
$= E\{0 + Z_{real}(\mathbf{u})Z_{real}(\mathbf{u})\} - E\{Z_{real}(\mathbf{u}) + 0\}E\{Z_{real}(\mathbf{u}) + 0\}$
$= Cov\{z_{real}(\mathbf{u}), z_{real}(\mathbf{u})\}$

As provas acima demonstram que erros independentes do teor não afetam a covariância quando $\mathbf{h} > 0$. Quando $\mathbf{h} = 0$, o covariância entre um ponto e ele mesmo é a sua variância $Var\{z_{obs}(\mathbf{u})\} = \sigma^2_{erro}(\mathbf{u}) = A_k^2$. O covariograma entre observações tem um efeito pepita $\lim_{\mathbf{h}\to 0} C_{obs}(\mathbf{h})$ composto de uma componente de microescala e uma ligada à componente de erro de amostragem. Apenas a componente em microescala faz parte do variograma do fenômeno real. O covariograma $C_{real}(\mathbf{h})$ pode ser inferido de $C_{obs}(\mathbf{h})$ a partir de:

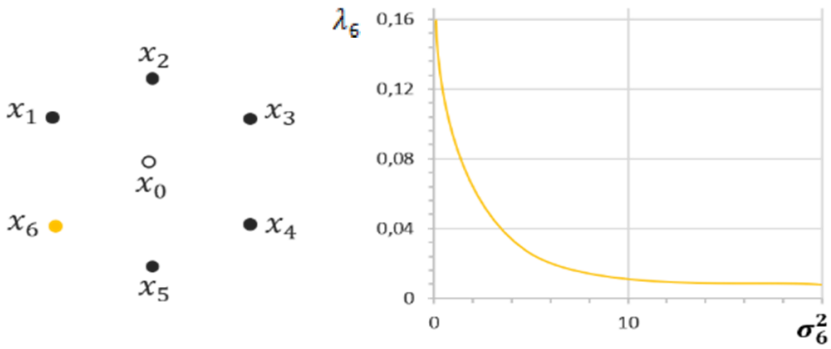
$$C_{real}(\mathbf{h}) = \begin{cases} C_{obs}(\mathbf{h}), & \mathbf{h} > 0 \\ C_{obs}(\mathbf{0}) - A_k^2, & \mathbf{h} = 0 \end{cases} \quad (13)$$

O modelo de covariância acima nos leva para a Equação 14, um sistema de krigagem cuja diagonal da matriz ($\mathbf{h} = 0$) é preenchida com a variância individual associada a cada observação. As células fora da diagonal são calculadas pelo covariograma da mesma forma que é feito entre OK e SK. O sistema de krigagem resultante é exatamente a KVME (DELHOMME 1976, p.95-99; WACKERNAGEL 2003).

$$\begin{pmatrix} C_{real}(\mathbf{0}) + A_1^2 & \cdots & C_{obs}(x_1 - x_n) \\ \vdots & \ddots & \vdots \\ C_{obs}(x_n - x_1) & \cdots & C_{real}(\mathbf{0}) + A_n^2 \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} C_{obs}(x_1 - x_0) \\ \vdots \\ C_{obs}(x_n - x_0) \end{pmatrix} \quad (14)$$

Como exemplo do funcionamento do KVME, Delhomme (1976) ilustra através da krigagem empregando seis amostras são distribuídas regularmente nos cantos de um hexágono (Figura 15). Cinco observações são livres de erros $\sigma^2_{erro}(\mathbf{u}) = 0$ enquanto $z_{6,obs}(\mathbf{u})$ é afetado por erros independentes do teor. O peso λ atribuído para estimar o valor no centro hexágono é o mesmo em todas amostras (1/6) quando o erro é nulo. Quando o erro aumenta, $\lambda_6(\mathbf{u})$ diminui, tendendo a zero quando o erro tende à variância *a priori* dos dados, o que equivale no sistema de krigagem à ausência de informação.

Figura 15 – Peso atribuído à X_6 em função o seu erro



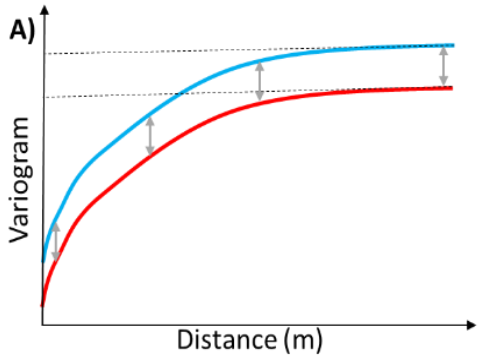
Medições sem erros X_i i=1,...,5 e a observação com erro X_6 são usados estimar o nó X_0. O gráfico à direita mostra o peso λ_6 atribuído a X_6 em função de seu erro associado. Modificado de Delhomme (1976).

Em geral, os erros e teores são correlacionados, e a suposição que Cov{$Z_{real}(\mathbf{u})$, $\varepsilon(\mathbf{u})$} = 0 é raramente atendida em dados reais. Portanto, a seguir será apresentado o desenvolvimento da equação para estimar o covariograma direto do fenômeno real através de observações afetadas por erros correlacionados com o teor.

3.5.1.2 Estimando o covariograma do fenômeno real - erro dependente do teor

Na presença de erros dependentes do teor, o covariograma $C_{real}(\mathbf{h})$ pode ser inferido a partir de $C_{obs}(\mathbf{h})$. As deduções e provas desse sistema são as mesmas daquelas apresentadas na Seção 4.2.2 Inferindo a estrutura de correlação real através de dados com erros correlacionados), sendo a equação da presente seção o caso especial em que não há erros compartilhados (Figura 16):

Figura 16 – Efeito no variograma de erros proporcionais ao teor



Efeito de erros independentes do teor entre o variograma observado (azul) e o real (vermelho).

Portanto, o tipo de erro é relevante para a inferência do covariograma real através do covariograma ajustado aos dados. O modelo pode ser inferido através das componentes de erro e do covariograma ajustado aos dados:

$$C_{real}(\mathbf{h}) = \begin{cases} C_{obs}(\mathbf{h}), & \mathbf{h} > 0 \\ C_{obs}(\mathbf{0}) - A_k^2 - z^2_{obs}(\boldsymbol{u}_i)C_k^2, & \mathbf{h} = 0 \end{cases} \quad (15)$$

Onde A_k^2 e C_k^2 correspondem a variância do erro de amostragem independente e dependente do teor, respectivamente. A Equação 13 é um caso especial de Equação 15 que ocorre quando $C_k = 0$. A covariância cruzada entre dois nós separados por uma distância $|\mathbf{h}|$ é inferida individualmente ao reescalonar $C_{real,(i,j)}$ em função do erro de amostragem associado a essas observações.

$$C_{obs,(i,j)}(\mathbf{h}) = \begin{cases} C_{real,(i,j)}(\mathbf{h}), & \mathbf{h}>0 \\ C_{real}(\mathbf{0}) + A^2_{(i,j)} + z^2_{obs}(\boldsymbol{u}_i)C^2_{(i,j)}, & \mathbf{h}=0 \end{cases} \quad (16)$$

Enquanto a Equação 13 leva a um sistema univariado que é idêntico ao KVME, a Equação 16 apresenta um modelo de covariância mais geral, que também é capaz de lidar com erros correlacionados ao teor. E aqui surge a maior vantagem do modelo de corregionalização apresentado: o sistema de krigagem para qualquer conjunto de dados com diferentes erros pode ser estimado diretamente de $C_{real,(i,j)}(\mathbf{h})$, desde que os erros associados à cada observação seja conhecidos, levando ao seguinte sistema de krigagem:

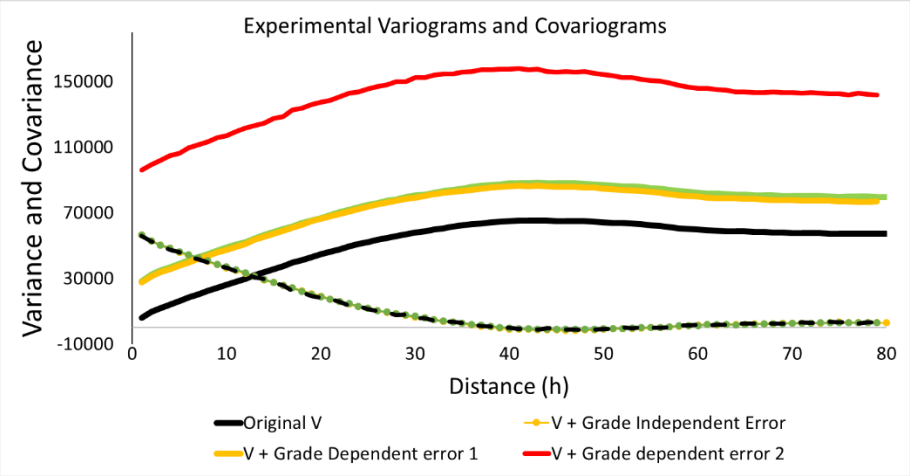
$$\begin{pmatrix} \mathrm{C}_{\mathrm{real}}(x_1 - x_1) + \mathrm{A}_1^2 + z^2{}_{obs} & (\boldsymbol{u}_1)\mathrm{C}_1^2 & \cdots & C(x_1 - x_n) \\ \vdots & & \ddots & \vdots \\ C(x_n - x_1) & & \cdots & \mathrm{C}_{\mathrm{real}}(x_n - x_n) + \mathrm{A}_j^2 + z^2{}_{obs} \quad (\boldsymbol{u}_j)\mathrm{C}_j^2 \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} C(x_1 - x_0) \\ \vdots \\ C(x_n - x_0) \end{pmatrix} \quad (17)$$

O método aumenta as possibilidades de simulação na presença de erros, assim como pode ser empregado por si só como um método mais geral do que o KVME para estimar modelos com krigagem. O desenvolvimento de métodos de krigagem na presença de erros não é escopo desse trabalho, sendo alguns pontos sobre essa aplicação sendo discutidos na Seção 5.2 RECOMENDAÇÕES PARA TRABALHOS FUTUROS. A seguir, um exemplo da influencia dos erros sob variogramas e covariogramas.

3.5.1.3 Exemplo 2 Estimando o variograma real

A Figura 17 mostra variograma e covariograma experimentais ajustados na direção norte-sul a observações obtidas dos valores iniciais da variável V (ppm) da versão exaustiva do banco de dados Walker Lake (Isaaks, Srivastava, 1989) perturbados por erros de diferentes tipos a magnitudes: Valores originais (linha preta), com erros independentes do teor A_k = 12 (linha verde), e os dados com erros proporcionais ao teor B_k = 0.63 (linha laranja), e B_k = 0.89 (linha vermelha). Todos os covariogramas são coincidentes, assim como os são os modelos ajustados aos dados com erros A_k = 12 e B_k = 0.63, linha vermelha e laranja, respectivamente. O desvio padrão das distribuições resultados são de 250, 291, 291 e 388 ppm, respectivamente.

Figura 17 – Relação entre covariograma e o variograma na presença de erros



Variogramas e covariogramas experimentais ajustados aos valores iniciais e a base de dados geradas através da adição de erros de amostragem (dependentes ou independentes do teor) aos valores iniciais.

Os resultados mostram que o variograma é consistente com o covariograma de referência, que corresponde ao fenômeno real. O desvio entre eles é dado pela variância do erro amostral adicionado. O covariograma é insensível aos tipos de erros adicionais em todos os intervalos onde h>**0**.

3.5.2 Inferindo a distribuição do fenômeno real

A distribuição ajustada a observações combina a distribuição real com a dos erros e, portanto, tem uma variância maior que aquela do fenômeno real. O problema inverso infere a distribuição real através do modelo de erros e da distribuição das observações. É inverso porque ele lida com problema começando pelos resultados para inferir a causa, nesse caso, começando com a distribuição das observações para inferir a distribuição do fenômeno sem a influência de erros. A metodologia é resumida em duas etapas:

i. o algoritmo infere uma primeira distribuição do fenômeno real através dos métodos de correção de suporte de volume. A solução se baseia no fato de que o aumento do suporte reduz a variabilidade de forma semelhante ao que aconteceria com o aumento da precisão das observações. O fator de ajuste, necessário para a transformação através do modelo gaussiano discreto (MATHERON, 1976; ZAGAYEVSKIY; DEUTSCH, 2011), é inferido em função do erroo de amostragem médio e o esperado que seja a variância do fenômeno real:

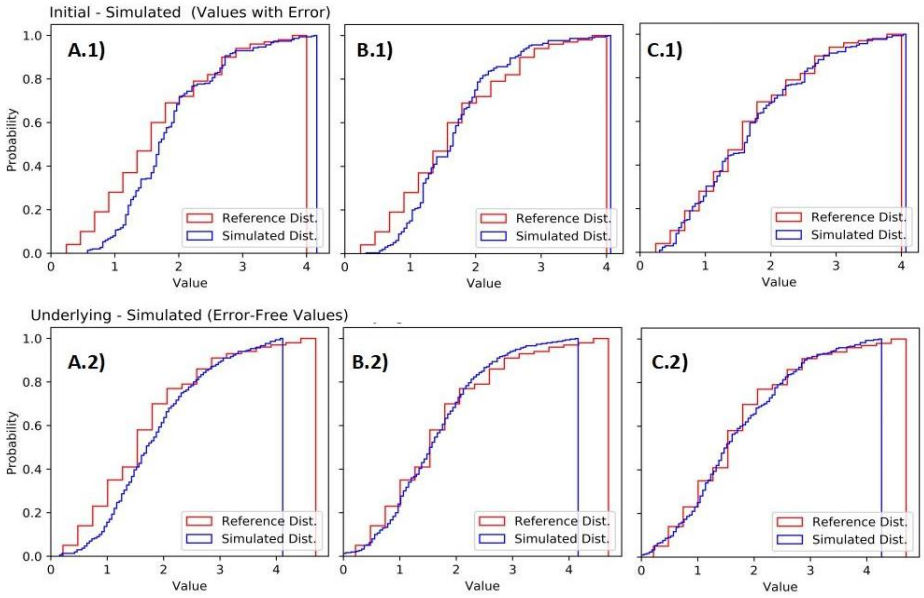
$$\text{Fator de ajuste} = 1 - \frac{\sigma^2_{y(\text{erro})(.)}}{\sigma^2_{Y_{\text{real}}(.)}} \quad (18)$$

ii. Realizações não-condicionais são simuladas em todas as posições amostradas. Cada valor gaussiano simulado $y^L_{k,real}(u_i)$ é perturbado por erros sorteados a partir de $N\{0, \sigma_{k,y\,(\text{erro})}(\mathbf{u})\}$. Como $Var\{Y_{k,obs}(\mathbf{.})\} - \sigma^2_{k,y\,(\text{erro})}(\mathbf{.}) = Var\{Y_{k,real}(\mathbf{.})\}$, a variância do fenômeno real $Y_{k,real}(\mathbf{.})$ é testada e atualizada até gerar um conjunto de valores gaussianos $y^L_{k,real}(\mathbf{u})$ (Figura 18a) que, quando perturbado por $N\{0,\sigma_{k,y(\text{erro})}(\mathbf{u})\}$ resulte em uma distribuição semelhante à distribuição $y_{k,obs}(\mathbf{u})$ (Figura 18b). A semelhança é objetivamente medida pela variância, distribuição/forma de histograma, assimetria, valores mínimos e máximos das duas distribuições.

A tabela de transformação das unidades originais para gaussianas é, então, usada para retrotransformar os valores gaussianos simulados para as unidades originais. No exemplo apresentando na Figura 18, as observações iniciais têm uma distribuição log-normal com média de 1.78, desvio padrão de 0.97, assimetria de 0.86, e valores mínimos e máximos de 0.21 e 4.69, respectivamente, e um erro absoluto é $\sigma_{erro}(\mathbf{.}) = 0.37$. A figura ilustra as etapas iniciais (A.1), intermediárias (B.1) e finais (C.1) (linha azul) do processo de otimização, em que o modelo inferido é testado e ajustado até coincidir com a distribuição inicial dos valores $z_{k,obs}(\mathbf{u})$.

A verdadeira distribuição é indisponível em problemas reais, e, portanto, a otimização é feita em relação à CDF ajustada aos dados. No entanto, a CDF do fenômeno real que a ser inferido é disponível nessa sintética como referência. A distribuição ajustada aos dados (linha vermelha, apresentada nas Figuras A.1, B.1 e C.1) é traçada nas mesmas três etapas da otimização em relação a verdadeira distribuição em A.2, B.2 e C.2.

Figura 18 - Comparação entre distribuição real e estimada



Comparação entre a mesma distribuição inicial (linha vermelha) e a distribuição simulada (linha azul) nas etapas: inicial (A.1), intermediária (B.1) e etapas finais (C.1) do processo de otimização. A associação real entre a distribuição real e a simulada antes da adição de erros é desconhecida em problemas reais, mas na mesma ordem, as figuras (A.2), (B.2) e (C.2) comparam a estimativa simulada $y^{L}_{k,real}(\mathbf{u})$ com a distribuição dos valores reais (linha azul).

Nessa seção, foi possível definir as estatísticas do fenômeno real que são necessárias para gerar simulações estocásticas. O modelo de corregionalização, que é definido automaticamente quando os erros associados às observações são conhecidos, sendo apenas necessária a modelagem do covariograma dos dados.

3.6 SIMULAÇÃO NA PRESENÇA DE ERROS

No primeiro passo do fluxo proposto, as observações devem ser substituídas por simulações assumidas como *hard data*. Para tal, a CoK é empregada para estimar a CCDF nas posições amostradas através do valor colocalizado, das observações da vizinhança e os *hard data* previamente simulados. A covariância espacial entre cada um desses pares é estimada através da Equação 16, que possibilita medir a degradação da informação entre um par de nós em função tanto da distância quanto do erro amostral associado a cada um. O erro associado a cada observação deve ser previamente inferido, enquanto que realizações de *hard data* são assumidas como isentas de erro.

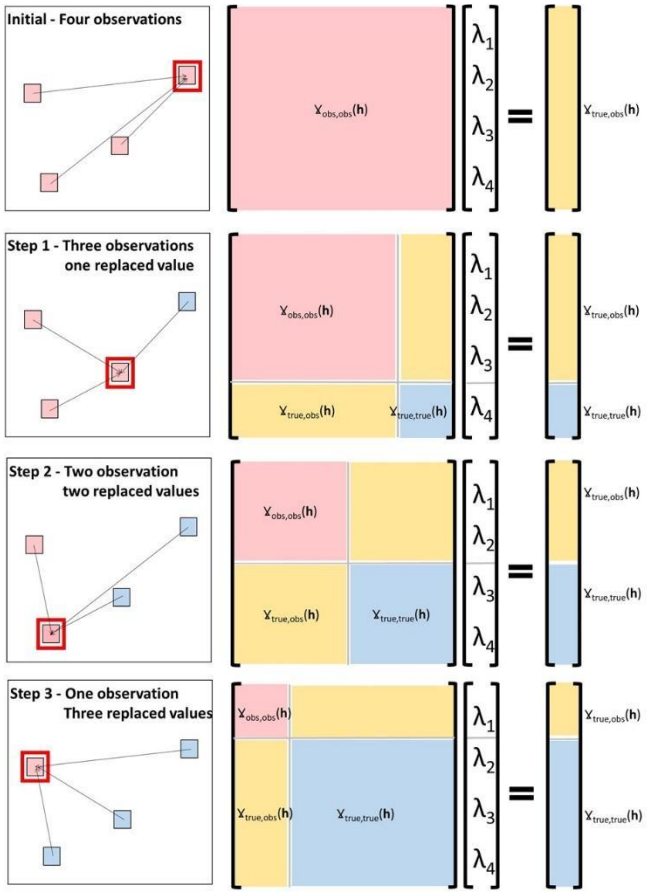
O formalismo original da CoK (MARECHAL, 1970) condiciona os pesos atribuídos a *hard data* a somarem um e soma zero aos pesos atribuídos aos *soft data*, o que traz consigo dois pontos de atenção

- condicionar a soma dos pesos dos *soft data* a 0 tende a limitar severamente a influência dos *soft data* (GOOVAERTS, 1998);
- condicionar a soma dos pesos dos *hard data* a 1 só é possível quando há ao menos uma observação desse tipo. A rotina de simulação de banco de dados faz com que em diversas vezes, haja apenas *soft data* em uma dada vizinhança.

Sob a condição assumida neste trabalho de que o erro de amostragem não é enviesado, $E\{\varepsilon(\mathbf{u})\} = 0$, a restrição acima apresentada é substituída por uma segunda em que o total dos pesos é condicionado a somar um. No APÊNDICE 1 – ESTIMANDO A CDF ATRAVÉS DE DADOS COM VIÉS), é apresentado outras duas situações possíveis do erro, que é a de erros com viés quando este é conhecido ou desconhecido.

A estimativa utiliza as observações colocalizadas, não-colocalizados e os valores previamente simulados. A CDF local é empregada para a simulação através de SGS ou outro método sequencial. A Figura 19 mostra quatro etapas na substituição das observações (indicados pela cor rosa) por *hard data* simulados (indicados pela cor azul).

Figura 19 – Esquema do sistema de CoK com a substituição das observações



Esquema do sistema de CoK para estimativa das CCDF na simulação do banco de dados. O sistema é composto pelos variogramas diretos e cruzados entre as observações (índice *obs)* e o fenômeno real (índice *true).* A diagonal (i = j) entre as observações é preenchida pelos erros associados aos dados individuais.

A variação entre realizações em um nó amostrado é controlada pelo erro associado à observação. Quanto maior o erro associado, menos redundantes são as observações não colocalizados na vizinhança e os valores previamente simulados. A Figura 20 mostra gráficos de dispersão entre os valores simulados $z^{L}_{real}(\mathbf{u})$ de uma realização e a observação inicial $z_{obs}(\mathbf{u})$ que ele a substituiu. O gráfico mostra a dispersão para dados com diferentes níveis de erros $\sigma^2_{erro}(\mathbf{u})$ e apresenta a correlação P_{expec} esperada em função do erro adicionado, e ρ_{meas} é a correlação efetivamente medida.

Figura 4 – Gráfico de dispersão entre observações e valores simulados

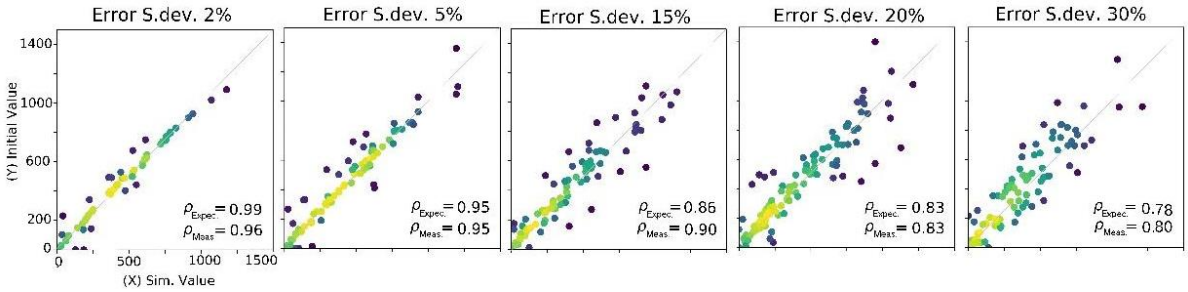


Gráfico de dispersão entre os valores originais (eixo Y) e os valores de uma realização (eixo X). A correlação P_{expec} é a correlação esperada em função do desvio-padrão do erro adicionado, e ρ_{meas} é a correlação efetivamente medida. As cores quentes indicam uma maior densidade de pontos.

Abaixo, é discutido mais detalhadamente a relação entre a variância de krigagem e o espaço de incerteza da simulação em um nó estimado e o comportamento da CoK com o modelo de corregionalização ajustado em função dos erros associados às amostras na vizinhança de cada nó a ser estimado tanto em casos normais quando nas situações limite.

A correlação ρ_{obs} entre uma observação e o fenômeno real na mesma posição é uma função do erro amostral associado à cada observação (Equação 23). Goovaerts (1997) mostra que o peso de co-krigagem atribuído a *soft data* λ_{soft} é proporcional à correlação ρ_{obs}. Quando a correlação é alta, as estimativas têm maior influência dos *soft data.* Essa influência é controlada pela atribuição dos pesos de krigagem λ_{hard} e λ_{soft}.

No formalismo da krigagem, o cálculo do peso é condicionado a gerar estimativas não enviesadas e minimizar o desvio quadrático $\{Z^*(\mathbf{u}) - Z(\mathbf{u})\}^2$, de forma a obtermos o BLUE (*best linear unbiased estimator*). A variância de krigagem resultante é utilizada para definir a variância da CCDF local centrada em $Z^*(\mathbf{u})$ pelos algoritmos de simulação sequencial. Considerando a simulação de novos valores em locais já amostrados **u**, temos duas situações limite:

1. $\sigma^2_{erro}(\mathbf{u}) = 0$ e $\rho_{obs} = 1$: todas as realizações em uma posição amostrada reproduzem o valor observado;
2. $\sigma^2_{erro}(\mathbf{u}) >= \sigma^2_{obs}(.)$: quando o erro associado a uma observação colocalizada é maior que a variância *a priori* dos dados, essa observação não adiciona qualquer informação na estimativa da CCDF, sendo, então, equivalente à ausência de dados nesta posição;

3. $0 < \sigma^2_{erro}(\mathbf{u}) < \sigma^2_{obs}(.)$: Nos casos em que o erro associado à observação é maior que zero, mas menor que a variância *a priori* dos dados, o modelo de correlação espacial pode ser atualizado em cada posição, assegurando a pesos de krigagem λ_{hard} e λ_{soft} que minimizem o erro de estimativa da CCDF. Essa solução é um avanço a COK, que, ao agrupar as observações em subgrupos, leva ao uso de ρ_{obs} médios que aumentam o erro de estimativa ao não considerar os erros individuais de cada amostra.

Portanto, a CoK empregando o modelo de covariância espacial apresentando na Equação 16, para estimar a CDF nas posições amostradas, garante valores consistentes tanto nas situações limites de ausência ou erros altos quanto a estimativa com menor erro de estimativa nos casos intermediários.

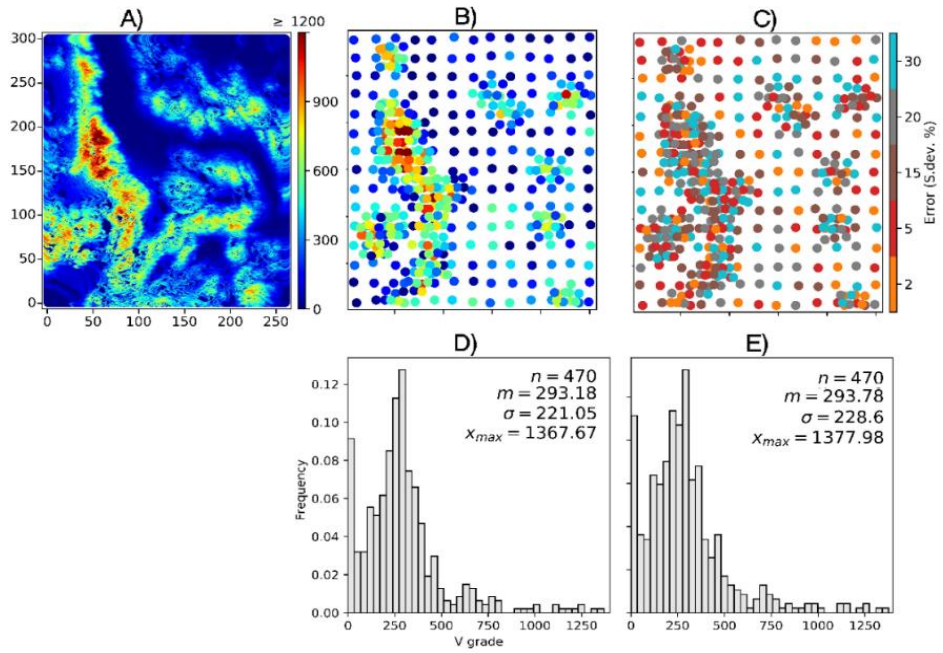
Na seção seguinte, será apresentando um exemplo da aplicação do método, em que são adicionados erros a um banco sintético e, então, comparado o desempenho do algoritmo apresentado com alternativas e o quanto elas se aproximam do modelo de referência, extensivamente amostrado e sem a presença de erros.

3.7 EXEMPLO 3 – SIMULANDO O FENÔMENO REAL ATRAVÉS DE AMOSTRAS COM ERROS

Este estudo de caso ilustra o método proposto. Ele usa o banco de dados Walker Lake, composto por 78.000 dados bidimensionais (Figura 21a) em uma malha regular 1x1 m ao longo de uma área de 280x300m (ISAAKS; SRIVASTAVA 1989). Do banco original, é derivado uma versão composta por 470 observações de V (ppm) em uma área, sendo 195 observações em uma malha pseudo-regular de 20x20m que imita uma malha de médio/longo prazo em um depósito mineral e 275 amostras em malha 5x5 m em áreas de alto teor, imitando uma malha de adensamento de *grade control* (Figura 21b). Os dados foram aleatoriamente divididos em cinco subgrupos Figura 21c) e os valores originais de originais de cada subconjunto foram perturbados por erros espacialmente heterogêneos, tirados de $N\{z_{obs}(\mathbf{u}), z_{obs}(\mathbf{u})*\varepsilon\}$, onde ε varia entre as amostras de 2%, 5%, 15%, 20% e 30%, imitando uma área de interesse que foi amostrada utilizando cinco métodos diferentes de amostragem, resultando em um único banco de dados com dados de diferentes qualidades.

A distribuição entre os erros é aleatória no espaço. Uma única tiragem de erros foi realizada, chegando a um único banco de dados inicial com erros, o que é a situação comum de casos reais.

Figura 21 – Mapas de Walker Lake dos valores originais e perturbados



A) Mapa exaustivo. B) localização das amostras. C) Mapa de erros de associados a cada amostra. Histogramas dos dados desagrupados D) sem. E) com erros.

A abordagem proposta utiliza o conjunto de dados redimensionado com erro (Figura 21e) para simular 100 realizações em uma malha 1 x 1 m que são, então, combinadas em blocos de 10 x 10 m. A Figura 22 mostra em detalhe as estatísticas de 5 bancos de dados dos 100 que foram simulados, considerando os valores iniciais com erros (de 2% a 30%) e o erro associado a cada um (Figura 21). É demonstrada a dispersão entre os valores de *hard data* simulados pareados com os valores iniciais, o coeficiente de correlação dessas relações varia entre 0.93 e 0.95, valores consistentes com o erro médio dos dados que é de 0.079. A dispersão entorno da linha x = y aumenta com o aumento dos teores, o que é esperado já que os erros são proporcionais a eles. Da mesma forma, as distribuições de *hard data* tiveram um desvio padrão de 0.92, demonstrado nos seus respectivos histogramas. Os variogramas estandardizados honram o alcance e variograma do modelo ajustado aos dados originais, sem a adição de erros (linhas verticais). A figura também mostra o mapa dos valores simulados nas posições das amostras inicias e, por fim, os mapas do grid 1x1m

que foram simulados, cada um condicionado a honrar um desses diferentes bancos de dados.

Figura 22 – Estatísticas dos bancos de dados simulados

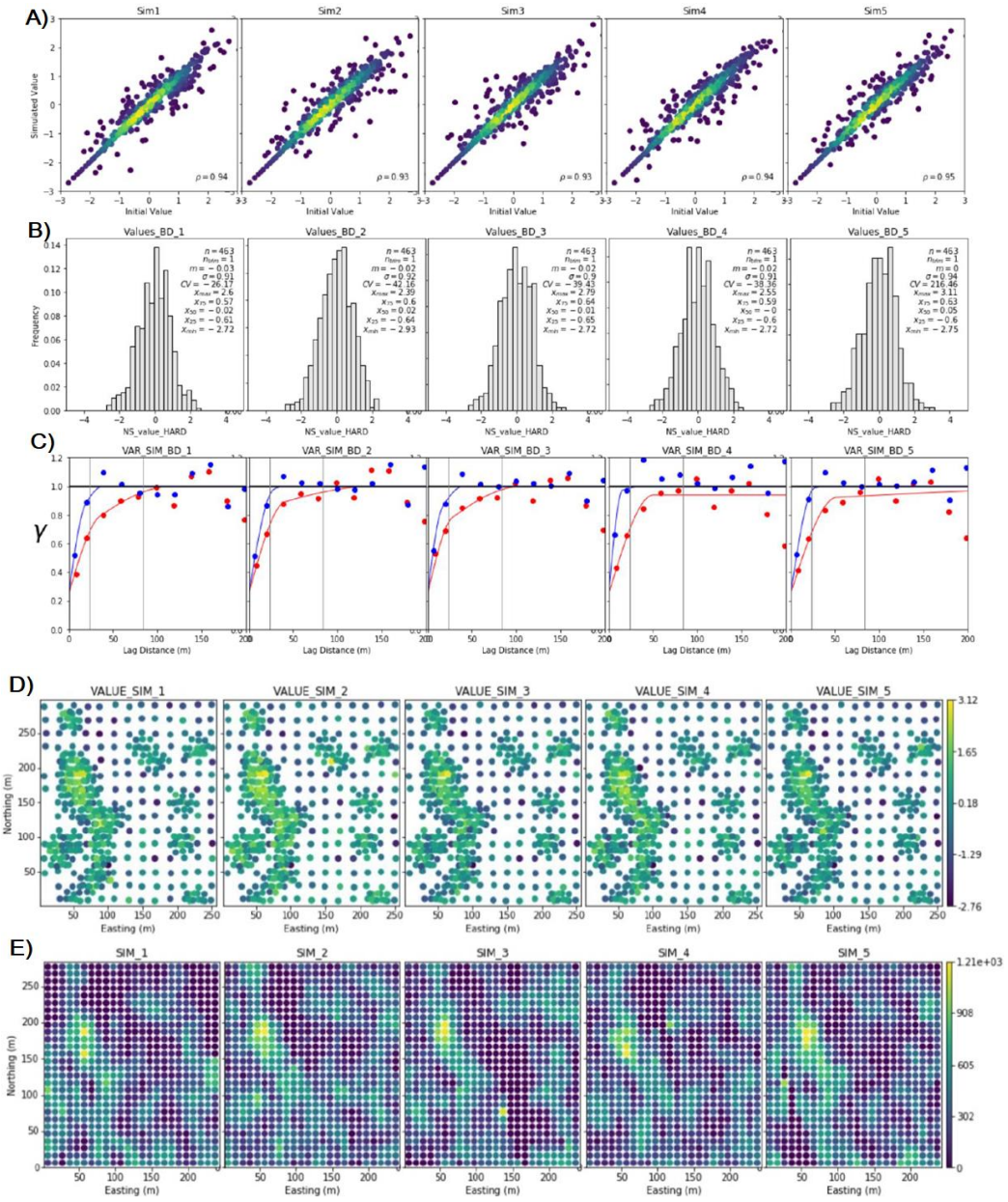
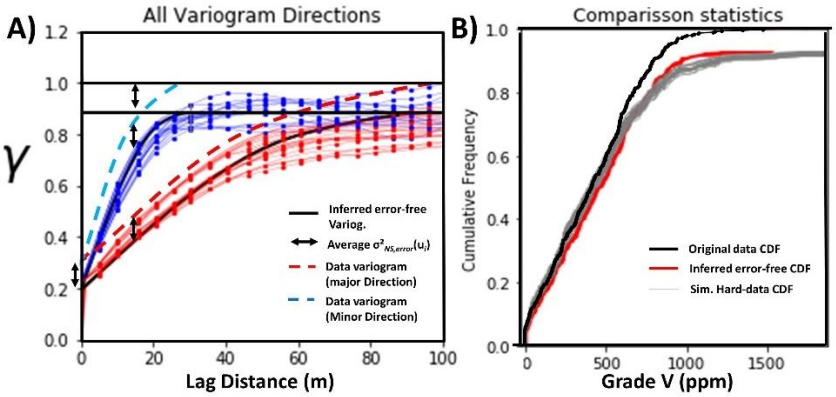


Gráfico de dispersão (A) entre valores simulados e originais, histogramas (B) e variogramas (C) de 5 bancos de dados simulados pelo método proposto, em que os valores iniciais foram substituídos por realizações de *hard data* simulados (D), os quais foram usados como *input* para simular o modelo regular de nós (E).

A Figura 23 apresenta simultaneamente todas as 100 realizações simuladas pelo método proposto, sendo os variogramas nas direções de maior (vermelho) e menor continuidade (azul) em relação ao variograma assumido como o do comportamento real do fenômeno (linha preta), as linhas finas vermelhas e azuis correspondem ao variograma

experimental das 100 realizações ao longo da maior e menor continuidade, respectivamente. Além disso, é apresentada a comparação entre a distribuição gaussiana dessas mesmas 100 realizações (linha cinza) em relação à distribuição desagrupada dos dados iniciais (linha vermelha). Ambas as figuras mostram uma reprodução dentro do esperado na reprodução das estatísticas condicionais inferidas do fenômeno real de interesse.

Figura 23 – Variograma e CDF das realizações simuladas pelo método proposto



A) Variograma ajustado aos dados Gaussianos na direção de maior (tracejado vermelho) e menor continuidade (tracejado azul) em relação ao modelo do comportamento do variograma sem a influência de erros (preto). As linhas finas vermelhas e azuis correspondem ao variograma experimental das 100 realizações ao longo da maior e menor continuidade, respectivamente. B) Distribuição das 100 realizações (cinza) em relação à distribuição do fenômeno real (vermelho) e do banco de dados inicial após o desagrupamento (preto).

Os resultados da metodologia proposta são comparados aos seguintes métodos:

- Os modelos gerados por SGS convencional, ignorando o erro de amostragem associado aos dados utilizados (Figura 21d), e prosseguindo por (i) transformando observações para unidades gaussianas, (ii) ajuste do variograma aos dados, (iii) simulação de 100 realizações e, (iv) retro transformação dos valores simulados para suas unidades originais.
- Modelo exaustivo, em que os valores do bloco são gerados pela média local do conjunto de dados exaustivos originais (malha 1 x 1 m) e assumidos como o valor real de cada bloco para avaliar a precisão das abordagens testadas.
- Modelo de referência, em que o banco de dados inicial sem erro é empregado para simular 100 realizações. O uso de dados sem erros torna correto assumir as realizações geradas como equiprováveis ao fenômeno real e seu espaço de incerteza como correto.

Os modelos são comparados quantitativamente em termos das estatísticas de cada realização, sua média (*E-type*) e a variância das realizações em cada bloco (Tabela 2). A Figura 1 (p. 16) é uma seção norte-sul ao longo x = 5 m entre realizações das abordagens propostas e convencionais. O método proposto tem uma menor variabilidade global, não honram os valores observados e tem um comportamento geral mais próximo do modelo de referência (linha preta).

Tabela 2 - Resumo do comparativo entre abordagens

		E-Type			Realizações individuais		
Modelo	**V ppm**	**S.Dev. Etype**	**CV**	**Coef. Corr. (Etype x exaustivo)**	**S.Dev. das 100 realiz.**	**CV**	**Média (Coef.Corr. x exaustivo)**
Mod. proposto	313.5	171.9	0.55	0.84	239	0.76	0.57
Mod. convenc.	349.8	185.8	0.53	0.85	323.2	0.92	0.61
Mod. referência	326.8	195	0.60	0.88	296	0.91	0.67
Mod. Exaustivo	295.4	221.5	0.75	1.00	221.5	0.75	1.00

Os modelos *E-type* são comparativos da estatística da média das 100 realizações. As "realizações individuais" mostram faz a média das estatísticas de cada realização individual em comparação com o conjunto de dados exaustivos. Os coeficientes de correlação são os valores da realização individual ou do *E-type* contra o modelo exaustivo. Ou seja, as estatísticas com o cabeçalho "E-type" é a estatística da média das realizações, enquanto que o "realizações individuais" é a média das estatísticas das realizações individuais.

É interessante considerar na análise da Tabela 2 que a redução da variabilidade em relação a variabilidade do banco de dados não pode ser entendida simplesmente como suavização, visto que o objetivo do método é a reprodução da variabilidade real do fenômeno, dada pelas estatísticas do modelo exaustivo.

O método proposto tem menor precisão do que o método convencional, indicado pela maior dispersão em torno da linha X = Y e menor correlação (Figura 24). A linha verde e vermelha de cada gráfico corresponde a regressão entre o método proposto e valores de referência (modelo de referência ou modelo exaustivo). A reta e as estatísticas em vermelho correspondem as regressões condicionadas a passarem pela origem, enquanto que a reta e estatísticas verdes correspondem a um modelo sem qualquer restrição imposta para o ajuste da reta.

Sendo a reta de regressão entre o valor de referência e o estimado, dada pela equação y = xa + b, a inclinação da reta, "b" é uma métrica comumente utilizada para diagnosticar o viés condicional global, onde um valor de "b" próximo de um é indica uma sobre estimativa em quantidade adequada. A figura 24 possibilita diversas conclusões sobre a comparação entre o método convencional e o proposto:

- No método convencional, a regressão é gerada condicionado que a = 0 (estatísticas e reta em vermelho, Figura 24B e Figura 24D). Nesse cenário, temos b > 1, indicando que o modelo tende a superestimar levemente a variância do modelo de referência, o que é esperado visto as realizações serem condicionadas a honrar um covariograma e distribuição com variâncias que combinam a variabilidade real com a do processo de amostragem, e portanto, superestimadas. É interessante lembrar que o E-type é uma estatística suavizada, sendo esperado que as realizações individuais tenham um "b" ainda maior do que 1;
- No método proposto, a regressão condicionada passar pela origem (estatísticas e reta em vermelho, Figura 24A e Figura 24C), a inclinação é 0.95 > b > 0.90, indicando uma suavização que é esperado para a estatística E-type;

Ao gerar retas de regressão onde o parâmetro "a" não é mais condicionado a ser zero, o seu valor é um indicador o viés nas faixas de baixo teor. Tanto comparando com o modelo de referência quanto ao modelo exaustivo, o método proposto apresenta um viés em baixos teores menor do que o método convencional. Ou seja, a suavização do método convencional tende a superestimar os baixos teores.

Figura 24 – Dispersão entre os valores das abordagens e dos valores de referência

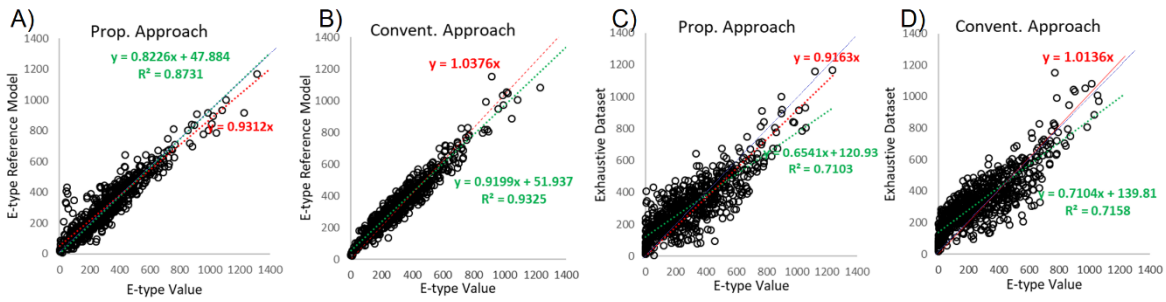
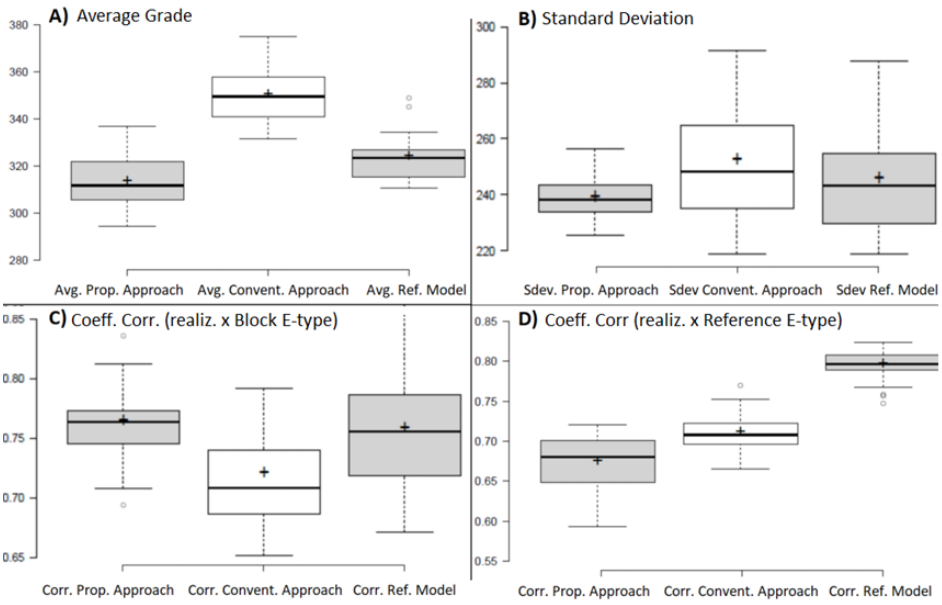


Gráfico de dispersão entre o *E-Type* do modelo de referência e o A) método proposto ou B) método convencional. Comparativo do modelo exaustivo e o C) método proposto ou D) método convencional. As estatísticas e retas y = xb + a em verde e vermelho correspondem, respectivamente, a regressão ajustada com e sem a restrição de que a = 0.

Outra análise para o entendimento da relação entre o método proposto e o convencional é o coeficiente de variação (CV) e o coeficiente de correlação entre o *E-type*, onde a abordagem proposta mostra uma melhor reprodução das características do fenômeno real (Figura 25).

As realizações de referência e do modelo proposto apresentam teores médios muito semelhantes, enquanto as realizações convencionais apresentam maior média, resultante de sua maior sensibilidade aos valores extremos/*outliers*, já que não foi utilizado etapa de *capping* (Figura 25a) As realizações simuladas pela abordagem proposta têm menos variância entre suas realizações individuais porque são limitadas para honrar o erro de amostragem associado a cada observação em vez de honrar o valor exatamente observado (Figura 25b). Na Figura 25C e Figura 25D, o coeficiente de correlação entre blocos simulados e o *E-type* do modelo ou entre os blocos simulados e a média do bloco real (média dos dados exaustivos), respectivamente. Esses resultados indicam que o espalhamento das realizações em torno do *E-type* é semelhante entre o modelo de referência e o proposto.

Figura 25 – *Boxplot* dos valores das abordagens e de referência



Boxplot de realizações simuladas pelo método proposto, abordagem convencional e modelo de referência. A) teor médio de V, B) desvio padrão do teor V, C) coeficiente de correlação de cada bloco nas realizações e o E-type médio de todas as realizações; D) correlação dos valores de cada realização e do E-type do modelo de referência.

As estatísticas e gráficos acima apresentam um comportamento global dos métodos analisados. A Figura 26 apresenta um método gráfico para comparar percentis de 5% a 95% (eixo x) com o seu correspondente teor (eixo y). Ao plotar o ponto (x, y) de diferentes modelos, uma maior semelhança entre duas distribuições em um percentil é indicada pela menor distancia vertical entre tais pontos.

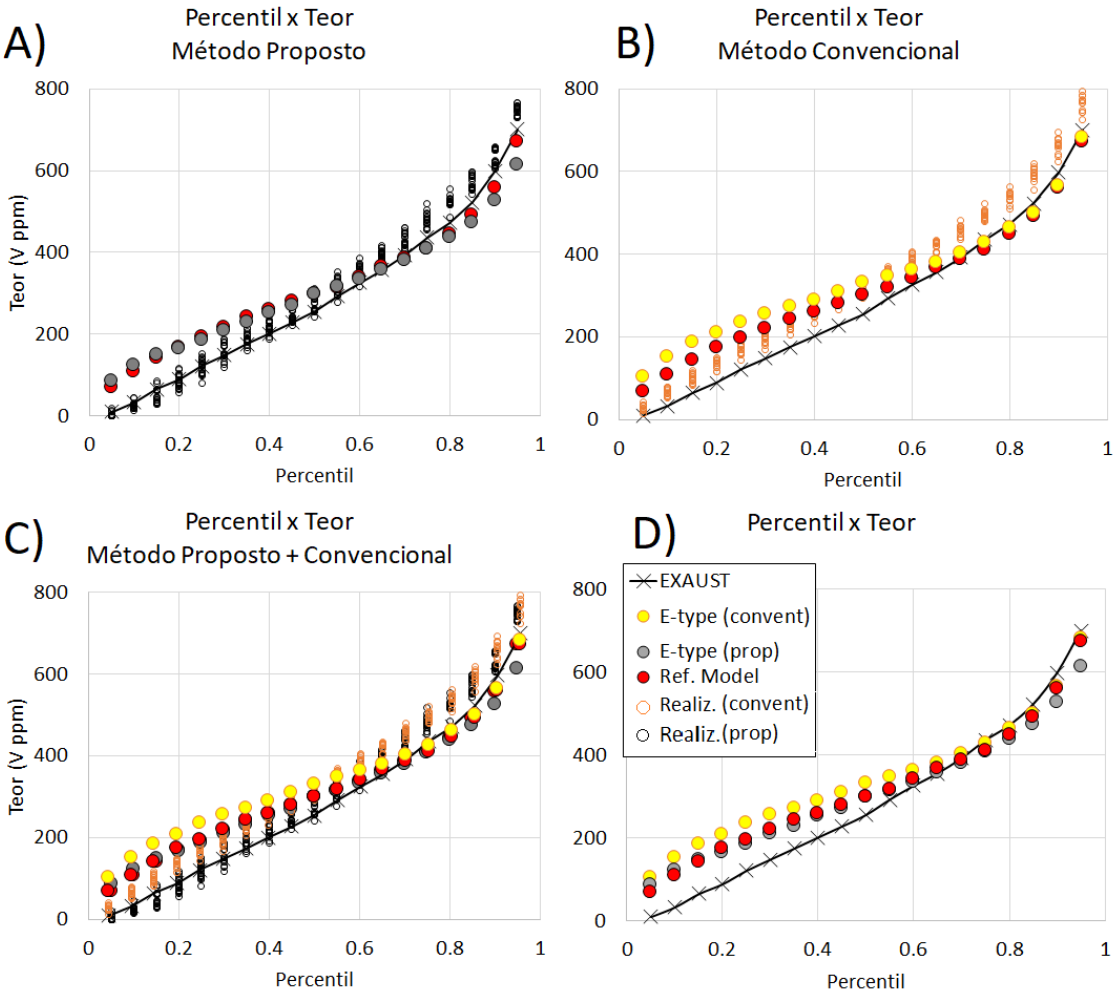
A Figura 26 apresenta diferentes combinações dos resultados obtidos, como forma de facilitar a comparação das realizações e o E-type do método proposto (Figura 26a), modelo convencional (Figura 26b) com o modelo exaustivo e modelo de referência. A Figura 26c apresenta todos esses dados combinados e, por último a Figura 26d apresenta apenas os E-type dos modelos e o modelo exaustivo.

Na Figura 26a, as realizações do método proposto (círculos pretos vazios) se distribuem entorno, ou próximo, aos valores do modelo exaustivo (reta contínua), sendo este assumida como o valor real do processo V. Isso mostra a capacidade do método proposto de gerar realizações que honram a distribuição do fenômeno de real, mesmo que na presença de dados com erros. A relação entre teor e percentil do E-type do modelo proposto (círculos cheios de cor cinza) indicam a suavização esperada ao fazer a média das realizações, mas ainda sim, eles se mantem extremamente próximos do modelo E-type de referência (círculos cheios de cor vermelha).

Na Figura 26b, as realizações do método convencional (círculos laranja vazios) se mantem acima do modelo exaustivo (reta continua), indicando um viés sistemático ao longo de todos percentis. A relação entre teor e percentil do modelo convencional (círculos cheios de cor laranja) indica viés, do percentil 5% a até o 75%, em relação ao E-type do modelo de referência (círculos cheios de cor vermelha)

As Figuras 26c e 26d possibilitam analisar o comportamento dos dois modelos em relação ao modelo exaustivo e ao de referência, mostrando que tanto globalmente quanto em diferentes percentis o método proposto apresenta uma melhor reprodução da distribuição do fenômeno de interesse.

Figura 26– *Comparação entre Percentis de teores (v ppm) entre modelos*



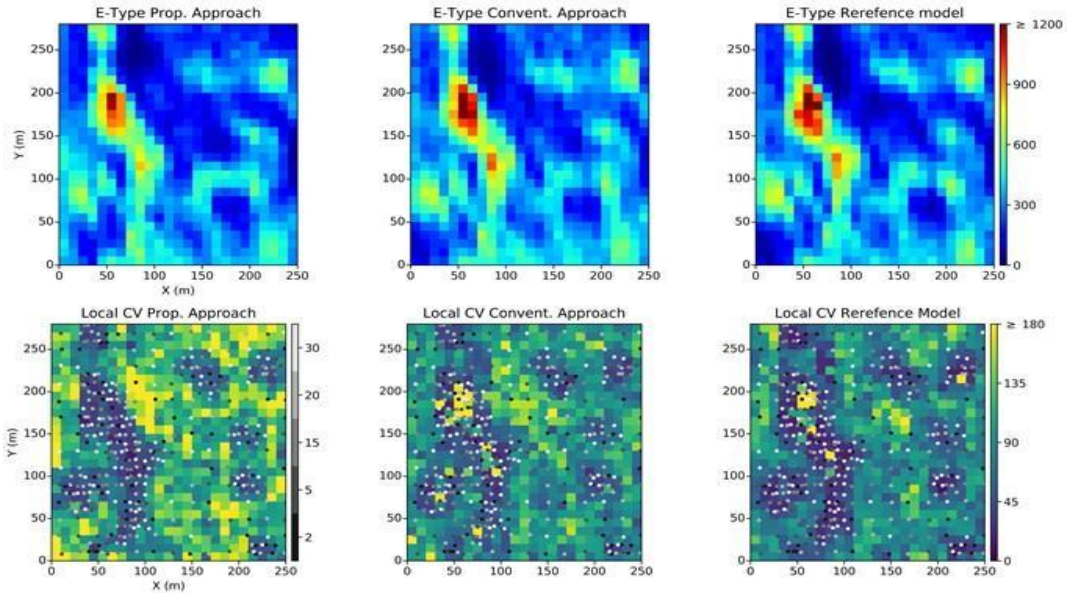
Comparativo dos percentis de 5% a 95% (eixo x) com o seu correspondente teor (eixo y) para as realizações individuais e modelo E-type do método A) proposto, B) convencional, C) ambos combinados e para facilidade de leitura, D) apenas os E-type. Em todos, os métodos são comparados como a distribuição dos dados exaustivos e o E-type do modelo de referência.

Os mapas de *E-type* (Figura 27) são muito semelhantes, o que é corroborado por sua correlação com os dados exaustivos (Tabela 2). Os mapas CV da abordagem convencional e o modelo de referência ignoram o erro amostral e são controlados pela densidade de dados e variabilidade dos valores locais. Pequenas diferenças entre eles vêm do uso de diferentes modelos de variogramas e as observações sem e com erros, respectivamente, a Figura 27d e Figura 27e.

O mapa de CV da abordagem proposta é influenciado pela densidade de dados locais (os valores mais baixos de CV estão em áreas densamente amostradas) e pelo erro amostral de cada observação. As realizações têm menor desvio padrão devido a serem condicionadas à variogramas e histogramas com menor variabilidade (Figura 25B). A variabilidade dentro

das realizações do método proposto é maior do que a da abordagem convencional, pois são condicionadas a bancos de dados diferentes. A variabilidade dentro das realizações é amplificada em áreas menos amostradas e/ou em que as observações têm grandes erros.

Figura 27 – *E-type* e mapas de coeficiente de variação (CV)



E-type e mapas de coeficiente de variação (CV) para a abordagem propostas e a convencional utilizando dados com erros. O modelo de referência é simulado empregando os dados sem a adição de erros.

Os resultados apresentados são condizentes com a teoria, em que a incerteza e valores do modelo gerados pelo método proposto são função: do erro associado à observação, da quantidade de dados e da distância da observação ao nó a ser simulado - o que são características condizentes com o esperado na realidade. Os resultados da abordagem proposta são mais próximos da variância, distribuição e do teor global de referência. Eles também mostram melhor reprodução da incerteza local e global. A precisão local de ambos os métodos é equivalente.

4 SIMULAÇÃO NA PRESENÇA DE ERROS: CASO MULTIVARIADO

Esta seção estende a teoria apresentada nas seções anteriores para o caso multivariado, em que cada amostra pode ser analisada para mais de 50 variáveis. As informações de interesse não são apenas os valores individuais de cada observação, mas também a estrutura de correlação entre eles. Portanto, através das observações e seus erros, é necessário inferir o variograma e o histograma de cada fenômeno e, também, a estrutura de correlação entre as variáveis.

A ocorrência de erros compartilhados está relacionada, principalmente, a variáveis associadas a minerais que são afetados de forma semelhante em um determinado processo de segregação durante a delimitação, extração, preparação ou etapas de subamostragem, gerando erros correlacionados. O entendimento de erros correlacionados entre as observações de diferentes variáveis é fundamental para a correta inferência do fenômeno real. Detalhado levantamento bibliográfico não encontrou nenhuma discussão na literatura geoestatística sobre observações que são afetadas por erros compartilhados.

4.1 ESTENDENDO O MODELO DE ERROS PARA O CASO MULTIVARIADO

No caso multivariado, o erro associado às observações de diferentes variáveis pode ser correlacionado. Na presença de erros que sejam independentes entre variáveis, quão maior o erro, menor a correlação medida através das observações. Na presença de erros correlacionados, a correlação e covariância medida pelas observações pode superestimar a associação real entre os fenômenos de interesse. Ignorar a proporção de erros compartilhados e assumir as estatísticas observadas como corretas pode induzir a estimativas erradas.

No Apêndice 2 (estimando as componentes proporcionais e independentes de erro), são apresentadas discussões sobre soluções para a estimativa de cada componente dos erros no caso multivariado. A hipótese da ocorrência de erros compartilhados é plausível e matematicamente desenvolvida nesta seção, mas o tema carece de mais aprofundamentos em relação a rotinas laboratoriais para determinar as suas componentes.

O erro correlacionado é o produto de fontes de erro que afetem de forma semelhante mais de uma variável. Uma amostra incorretamente coletada introduz a falta de equiprobabilidade na amostragem, fazendo com que haja um viés na chance de amostrar

um dado elemento. A ocorrência de erros compartilhados está relacionada, principalmente, a variáveis associadas a minerais que são afetados de forma semelhante em um determinado processo de segregação durante a delimitação, extração, preparação ou subamostragem.

A Figura 28 ilustra quatro níveis diferentes de associação física entre duas variáveis. Na Figura 28A ambas as variáveis estão associadas à mesma estrutura mineral e, portanto, seus erros de amostragem são compartilhados em qualquer escala e granulometria. A Figura 28B e a Figura 28C ilustram níveis de associação intermediários, e a Figura 28D, o caso que há independência física entre os minerais que portam os elementos de interesse.

Figura 28 - Quatro níveis diferentes de associação entre os elementos

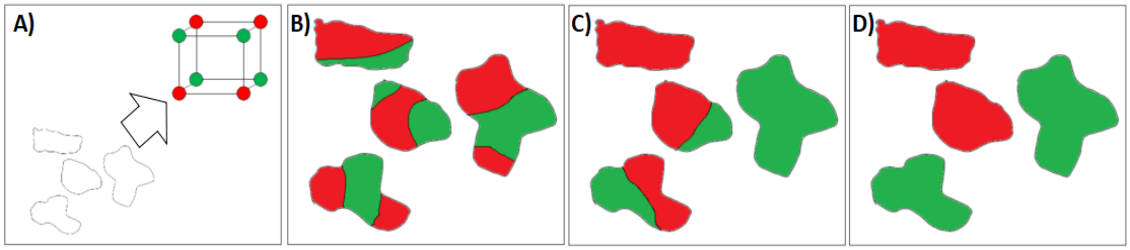


Ilustração de quatro níveis diferentes de associação entre os elementos A (vermelho) e B (verde). Na Figura A) ambos os elementos compartilham a mesma estrutura de montagem mineral e, portanto, seus erros de amostragem são compartilhados em qualquer escala e granulometria. Figura D) ilustra a independência completa entre os minerais e, em seguida, o erro compartilhado pode ser causado apenas por propriedades semelhantes que levam à segregação semelhante. Os números B) e C) ilustram associações intermediárias.

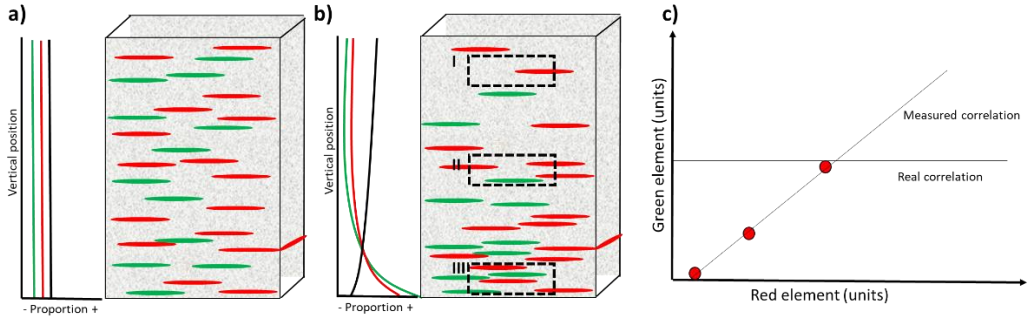
Exemplos de propriedades que podem introduzir viés simultaneamente a duas ou mais variáveis: (i) forma e densidade; (ii) propriedades magnéticas; (iii) propriedades eletrostáticas de constituintes liberados, como flocos minúsculos de biotita; (iv) teor de umidade e/ou capacidade de absorção de água; (v) minerais aderindo de forma diferente às paredes de uma caixa de armazenamento ou com ângulo diferente de repouso etc. (PITARD, 1993).

O exemplo a seguir ilustra uma correlação observada das amostras tomadas de um lote altamente segregado que mostra como a associação medida entre duas variáveis pode ser diferente da real.

4.2 EXEMPLO 4 - OCORRÊNCIA DE ERROS COMPARTILHADOS–CONCEITUAL

Neste exemplo, analisamos a relação entre dois minerais com forma de flocos (vermelho e verde) que sejam, em seu estado de homogeneidade inicial, espacialmente independentes um do outro (Figura 29a). As três áreas tracejadas (I, II e III) são amostras diretamente coletadas do lote não homogeneizado (Figura 29b).

Figura 29 - Esquema do efeito de segregação na associação medida



Lote composto de minerais de ganga (cinza) e dois tipos de minerais micáceos em a) sua organização original *in situ,* aleatoriamente distribuídos; b) o lote altamente segregado e com três amostras delimitadas (quadrados tracejados); c) uma regressão ajustada à proporção dos dois minerais nas três amostras delimitadas.

Os minerais em cada amostra estão simultaneamente acima ou abaixo da proporção global do lote, o processo de segregação cria uma correlação falsa. Neste caso, avaliar o erro que surge do processo de segregação e removê-lo de cada variável não é suficiente para estimar a correlação real entre as variáveis. Só é possível fazê-la se a presença de um erro compartilhado fosse levada em conta.

4.2.1 Modelo matemático para erros compartilhados

No caso multivariado, há m = 1,., M componentes de erros, onde o erro total M_{total} é a combinação dos erros proporcionais M_{rel} e os absolutos M_{abs}. A tabela de valores $\mathbf{ind}_{k,m}$ (Figura 30) indica quais variáveis *K* são afetadas (marcadas com número 1) pela componente M de uma dada fonte de erros. Quando $\mathbf{ind}_{k,m} = 1$ para apenas uma variável *K*, apenas esta é afetada por essa fonte de erro. Quando as linhas correspondentes a duas ou mais variáveis *K* tem em uma dada coluna M o valor de $\mathbf{ind}_{k,m} = 1$, esta componente de erro é compartilhada entre elas (Figura 30).

Figura 30 - Esquema da matriz das componentes associadas a cada elemento

$$Ind_{k,m} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \dots \\ K \end{array} \begin{array}{c} \begin{array}{ccccc|ccccc} m=1 & m=2 & m=3 & m=\dots & m=M_{rel.} & m=(M_{rel.}+1) & m=(M_{rel.}+2) & m=(M_{rel.}+3) & m=\dots & m=M_{total} \end{array} \\ \left(\begin{array}{ccccc|ccccc} 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \end{array} \right) \end{array}$$

Esquema da matriz **Ind**$_{k,m}$ que mostra quais variáveis *K* são afetadas por quais componentes do M$_{total}$, divididos em erros proporcionais ao teor M$_{rel}$ (retângulo verde) e erros não correlacionados com M$_{abs}$ (retângulo amarelo).

A magnitude do erro de uma dada fonte M é o produto de **ind**$_{k,m}$ = 1 multiplicada pela magnitude da **B**$_m$, conforme presente na matriz **b**$_{k,m}$ (Figura 30 e Figura 31). Por exemplo, dada uma componente de erro m = 1, com magnitude **B**$_1$ = 2 e que são compartilhadas pelas variáveis *K* = 1 e 3, temos: **b**$_{1,1}$ = 1***B**$_1$ = 2, **b**$_{2,1}$ = 0***B**$_1$ = 2, **b**$_{3,1}$ = 1***B**$_1$ = 0. A soma dos quadrados de todos os componentes de uma única linha é a variância total do erro associada a variável *K*, portanto, no exemplo apresentando, a contribuição de *K* = 1, 2 e 3 é de $2^2 + 0^2 + 2^2 = 8$

Figura 31 - Exemplo da matriz das componentes associadas a cada elemento

$$b_{k,m} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \dots \\ K \end{array} \begin{array}{c} \begin{array}{cccccccccc} m=1 & m=2 & m=3 & m=\dots & m=M_{rel.} & m=(M_{rel.}+1) & m=(M_{rel.}+2) & m=(M_{rel.}+3) & m=\dots & m=M_{total} \end{array} \\ \left(\begin{array}{cccccccccc} 1.B_1 & 0.B_2 & 1.B_3 & 0.B_{\dots} & 0.B_M & 1.B_{m+1} & 1.B_{m+2} & 0.B_{m+3} & 1.B_{\dots} & 0.B_{2M} \\ 0.B_1 & 1.B_2 & 0.B_3 & 0.B_{\dots} & 1.B_M & 0.B_{m+1} & 1.B_{m+2} & 0.B_{m+3} & 0.B_{\dots} & 1.B_{2M} \\ 1.B_1 & 0.B_2 & 0.B_3 & 1.B_{\dots} & 0.B_M & 0.B_{m+1} & 1.B_{m+2} & 0.B_{m+3} & 1.B_{\dots} & 0.B_{2M} \\ 0.B_1 & 0.B_2 & 0.B_3 & 1.B_{\dots} & 1.B_M & 1.B_{m+1} & 0.B_{m+2} & 0.B_{m+3} & 0.B_{\dots} & 1.B_{2M} \\ 1.B_1 & 1.B_2 & 0.B_3 & 0.B_{\dots} & 1.B_M & 0.B_{m+1} & 0.B_{m+2} & 1.B_{m+3} & 1.B_{\dots} & 0.B_{2M} \end{array} \right) \end{array}$$

Exemplo da matriz **b**$_{k,m}$ onde cada célula é o produto de um indicador **ind**$_{k,m}$ com valor de 0 ou 1 que indica quais variáveis K são afetadas por cada componente do erro M a magnitude B$_m$ relacionados a cada um desses esses indicadores. Os componentes são divididos em componentes relativos ao teor M$_{rel}$, e erros independentes M$_{abs}$.

O modelo de erros (Equação 1) é estendido a um modelo que lida com múltiplos componentes de erro independentes e compartilhados (*m* = 1,...,*M*). O modelo considera que cada observação é afetada por erros que variam devido ao (i) tipo de variável *(k = 1,...,K)*, (ii) localização **u**, e (iii) o tipo de erro, sendo eles correlacionados ou independentes do teor.

$$z_{k,obs}(\mathbf{u}) = z_{k,real}(\mathbf{u}) + \underbrace{z_{k,real}(\mathbf{u}) \sum_{m=1}^{M_{rel.}} b_{k,m} X_m(\mathbf{u})}_{\text{Erros relativos}} + \underbrace{\sum_{m=M_{rel.}+1}^{M_{total}} b_{k,m} X_m(\mathbf{u})}_{\text{Erros Abs.}} \quad (19)$$

O desvio entre $z_{k,real}(\mathbf{u})$ e os valores $z_{1,obs}(\mathbf{u})$ ou $z_{2,obs}(\mathbf{u})$ resulta de erros absolutos e proporcionais ao teor. Quando uma determinada fonte de erro é compartilhada entre as

observações de variáveis *K,* o mesmo erro vindo da componente $X_m(\mathbf{u})\mathbf{b}_{k,m}$ é compartilhado entre eles. Todos os componentes $\mathbf{b}_{k,m,}$ de qualquer conjunto de variáveis, *K* podem ser agrupados nas seguintes componentes: erro proporcional ao teor e compartilhado (D, e não compartilhado (C_k), bem como erros independentes são compartilhados (B) e não compartilhados (A_k):

$$z_{k,obs} = z_{k,real} + \underbrace{A_k X_1 + BX_3}_{\text{erros indep. do teor}} + \underbrace{z_{real}\{C_k X_2 + DX_4\}}_{\text{Erro prop. ao teor}} \quad (20)$$

A Equação 20 é uma generalização dos modelos anteriores, abrangendo casos como o de observações afetadas apenas por erros aditivos (C_k = D = 0) ou erros multiplicativos (A_k = B = 0) ou combinações entre eles. O modelo da Equação 1 é um caso específico em que o erro entre variáveis é independente um do outro (A_k = C_k = 0).

4.2.2 Inferindo a estrutura de correlação real através de dados com erros correlacionados

Depois de definir o modelo generalizado de erros, agora discutimos como inferir as estatísticas do fenômeno real através das observações. A variância da Equação 2 combina a amostragem e erros analíticos mais a variância do fenômeno real:

$$\text{Var}\{Z_{k,obs}\} = A_k^2 + B^2 + \text{Var}\{Z_{k,real}\}\left(1 + C_k^2 + D^2\right) \quad (21)$$

A $\text{Cov}\{Z_{1,real}, Z_{2,real}\}$ pode ser estimada a partir de observações em um determinado domínio, sendo relevante para o entendimento do caso multivariado. Utilizando a propriedade de bilinearidade (Equação 12) e a mesma notação simplificada para a covariância, chegamos à seguinte equivalência entre a covariância real e a ajustadas a observações com erros (Equação 22):

Prova 2:

$\text{Cov}\{Z_{1,true}, Z_{2,true}\} = [Z_{1,obs} - A_1X_{1,1} - BX_3 - Z_{1,true}C_1X_{1,2} - Z_{1,true}DX_4, Z_{2,obs} - A_2X_{2,1} - BX_3 - Z_{2,true}C_2X_{2,2} - Z_{2,true}DX_4]$

$= [Z_{1,obs}, Z_{2,obs}] - [Z_{1,obs}, A_2X_{2,1}] - [Z_{1,obs}, BX_3] - [Z_{1,obs}, Z_{2,true}C_2X_{2,2}] - [Z_{1,obs}, Z_{2,true}DX_4]$
$-[A_1X_{1,1}, Z_{2,obs}] + [A_1X_{1,1}, A_2X_{2,1}] + [A_1X_{1,1}, BX_3] + [A_1X_{1,1}, Z_{2,true}C_2X_{2,2}] +$
$[A_1X_{1,1}, Z_{2,true}DX_4]$
$-[BX_3, Z_{2,obs}] + [BX_3, A_2X_{2,1}] + [BX_3, BX_3] + [BX_3, Z_{2,true}C_2X_{2,2}] + [BX_3, Z_{2,true}DX_4]$
$-[Z_{1,true}C_1X_{1,2}, Z_{2,obs}] + [Z_{1,true}C_1X_{1,2}, A_2X_{2,1}] + [Z_{1,true}C_1\ X_{1,2}, BX_3] +$
$[Z_{1,true}C_1X_{1,2}, Z_{2,true}C_2X_{2,2}] + [Z_{1,true}C_1X_{1,2}, Z_{2,true}DX_4] - [Z_{1,true}DX_4, Z_{2,obs}] + [Z_{1,true}DX_4, A_2X_{2,1}] +$
$[Z_{1,true}DX_4, BX_3] + [Z_{1,true}DX_4, Z_{2,true}C_2X_{2,2}] + [Z_{1,true}DX_4, Z_{2,true}DX_4]$

Isolando as constantes baseado em Cov{aX,Y} = aCov{X,Y}:

$=[Z_{1,obs}, Z_{2,obs}]- A_2\ [Z_{1,obs}, X_{2,1}] - B[Z_{1,obs}, X_3] - C_2[Z_{1,obs}, Z_{2,true}X_{2,2}] - D\ [Z_{1,obs}, Z_{2,true}X_4]$
$-A_1\ [X_{1,1}, Z_{2,obs}]+ A_1A_2\ [X_{1,1}, X_{2,1}] +A_1B\ [X_{1,1}, X_3] +A_1C_2\ [X_{1,1}, Z_{2,true}X_{2,2}] + A_1D[X_{1,1}, Z_{2,true}X_4]$
$-B[X_3, Z_{2,obs}]+ BA_2[X_3, X_{2,1}] + B^2[X_3, X_3] + BC_2\ [X_3, Z_{2,true}X_{2,2}] + BD[X_3, Z_{2,true}X_4]$
$-C_1\ [Z_{1,true}X_{1,2}, Z_{2,obs}]+ C_1\ A_2\ [Z_{1,true}X_{1,2}, X_{2,1}] + C_1\ B[Z_{1,true}\ X_{1,2}, X_3] +$
$C_1C_2[Z_{1,true}X_{1,2}, Z_{2,true}X_{2,2}] +C_1\ D\ [Z_{1,true}X_{1,2}, Z_{2,true}X_4]$
$- D\ [Z_{1,true}X_4, Z_{2,obs}]+ DA_2\ [Z_{1,true}X_4,\ X_{2,1}] + DB[Z_{1,true}X_4, X_3] + DC_2\ [Z_{1,true}X_4, Z_{2,true}X_{2,2}] +$
$D^2[Z_{1,true}X_4, Z_{2,true}X_4]$

Na próxima etapa, eliminamos todos locais onde Cov{X,Y} = 0

$=[Z_{1,obs}, Z_{2,obs}]-$ ~~$A_2\ [Z_{1,obs}, X_{2,1}]$~~ $-$ ~~$B\ [Z_{1,obs}, X_3]$~~ $- C_2\ [Z_{1,obs}, Z_{2,true}X_{2,2}] - D\ [Z_{1,obs}, Z_{2,true}X_4]$
~~$-A_1\ [X_{1,1}, Z_{2,obs}]+ A_1A_2\ [X_{1,1}, X_{2,1}] +A_1B\ [X_{1,1}, X_3] +A_1C_2\ [X_{1,1}, Z_{2,true}X_{2,2}] + A_1D[X_{1,1}, Z_{2,true}X_4]$~~
~~$-B[X_3, Z_{2,obs}]+ BA_2[X_3, X_{2,1}]$~~ $+ B^2[X_3, X_3]$ ~~$+ BC_2\ [X_3, Z_{2,true}X_{2,2}] + BD[X_3, Z_{2,true}X_4]$~~
$-C_1\ [Z_{1,true}X_{1,2}, Z_{2,obs}]$ ~~$+ C_1\ A_2\ [Z_{1,true}X_{1,2}, X_{2,1}] + C_1\ B[Z_{1,true}\ X_{1,2}, X_3]$~~ $+$
~~C_1~~$C_2[Z_{1,true}$~~$X_{1,2}$~~$, Z_{2,true}X_{2,2}] +C_1\ D\ [Z_{1,true}X_{1,2}, Z_{2,true}X_4]$
$- D\ [Z_{1,true}X_4, Z_{2,obs}]+$ ~~$DA_2\ [Z_{1,true}X_4,\ X_{2,1}] + DB[Z_{1,true}X_4, X_3]$~~ $+ DC_2\ [Z_{1,true}X_4, Z_{2,true}X_{2,2}] +$
$D^2[Z_{1,true}X_4, Z_{2,true}X_4]$

O que nos leva à:

$=[Z_{1,obs}, Z_{2,obs}]-$ ~~$C_2\ [Z_{1,obs}, Z_{2,true}X_{2,2}] - D\ [Z_{1,obs}, Z_{2,true}X_4]$~~ $+ B^2[X_3, X_3]$ ~~$-C_1\ [Z_{1,true}X_{1,2}, Z_{2,obs}] +$~~
~~$C_1C_2[Z_{1,true}X_{1,2}, Z_{2,true}X_{2,2}] +C_1\ D\ [Z_{1,true}X_{1,2}, Z_{2,true}X_4] - D\ [Z_{1,true}X_4, Z_{2,obs}] + DC_2$~~
~~$[Z_{1,true}X_4, Z_{2,true}X_{2,2}]$~~ $+ D^2[Z_{1,true}X_4, Z_{2,true}X_4]$

Tratando o último termo, temos que:

Cov{ab,ac}=E[a²bc]−E[ab]E[ac]=E[a²]E[b]E[c]−E[a]²E[b]E[c]:

$= D^2[X_4Z_{1,true}, X_4Z_{2,true}] = D^2\{E[X_4{}^2]E[Z_{1,true}]E[Z_{2,true}]-E[X_4]^2E[Z_{1,true}]E[Z_{2,true}]\}$

Sendo $E[Z_{true}] = E[Z_{obs}]$

$=D^2\{E[X_4{}^2]\ E[Z_{1,true}]E[Z_{2,true}] -E[X_4]^2\ E[Z_{1,true}]E[Z_{2,true}]\}$
$=D^2\ E[Z_{1,true}]E[Z_{2,true}]\{E[X_4{}^2] -E[X_4]^2\ \}$

onde $E[X_4{}^2] -E[X_4]^2 = [X_4] = Var\{X_4\} = 1$
$= D^2\ E[Z_{1,true}]E[Z_{2,true}]$

Finalmente chegamos à:

$[Z_{1,true}, Z_{2,true}] = [Z_{1,obs}, Z_{2,obs}] + B^2 + D^2\{E(Z_{1,obs})E(Z_{2,obs})\}$

Portanto, os termos aditivos não compartilhados C_k são cancelados e a covariância só é afetada por erros compartilhados entre as duas observações:

$$\mathrm{Cov}\{Z_{1,obs}, Z_{2,obs}\} = \mathrm{Cov}\{z_{1,\mathrm{real}}, z_{2,real}\} + B^2 + D^2\{E(Z_{1,obs})E(Z_{2,obs})\} \qquad (22)$$

A correlação entre fenômenos reais é a covariância entre duas variáveis livres de erros divididas pelo produto de seus desvios padrão. As Equações 21 e 22 levam a:

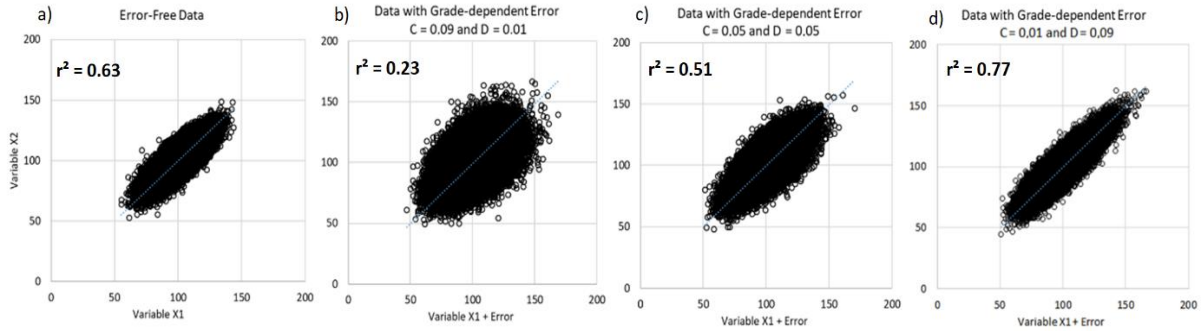
$$\rho\{Z_{1,real}, Z_{2,real}\} = \text{Cov}\{z_{1,obs}, z_{2,obs}\} - \quad (23)$$
$$B^2 - D^2\{E(Z_{1,obs})E(Z_{2,obs})\}\left(\frac{(1 + C_1^2 + D^2)(1 + C_2^2 + D^2)}{(\sigma_{1,obs} - A_1^2 - B^2)(\sigma_{2,obs} - A_2^2 - B^2)}\right)^2$$

A variância *a priori* do fenômeno real pode ser estimada pela diferença entre a variância observada e a do erro de amostragem $\sigma^2_{k,real} = \sigma^2_{k,obs} - \sigma^2_{k,erro}$. A estrutura de correlação do fenômeno real não pode ser medida diretamente, mas pode ser inferida a partir de observações se a natureza do erro for conhecida.

$$\rho\{Z_{1,real}(\mathbf{u}), Z_{2,real}(\mathbf{u})\} = \frac{\text{Cov}\{Z_{1,real}(\mathbf{u}), Z_{2,real}(\mathbf{u})\}}{\sigma_{Z_{1,real}(\mathbf{u})}\sigma_{Z_{2,real}(\mathbf{u})}} \quad (24)$$

A (Figura 32) mostra observações entre duas variáveis inicialmente sem erros com correlação $r^2 = 0.63$. As variáveis foram perturbadas por um erro total $C_k + D = 0.1$ onde D varia entre 0.01, 0.05 e 0.09. O r^2 após a adição do erro é respectivamente 0.23, 0.51 e 0.77. A correlação medida entre variáveis com erros compartilhados pode ser maior que a correlação entre os fenômenos reais dependendo da variância da componente de erro compartilhada.

Figura 32– Efeito dos diferentes tipos de erro na associação medida



A) gráfico de dispersão entre duas variáveis sem erros com uma correlação r^2 de 0.63. As variáveis foram perturbadas por um erro C + D = 0.1 com a proporção de D de 10, 50 e 90%. A correlação medida após a adição do erro é de 0.23, 0.51 e 0.77, respectivamente em b), c) d).

O exemplo a seguir mostra o impacto de erros compartilhados na medição das estatísticas e como a correlação medida entre duas variáveis pode ser diferente do real nesses casos.

4.2.3 Exemplo 5 ocorrências de erros compartilhados

Uma série de observações são simuladas para ilustrar como os quatro tipos de erro afetam a variância, a covariância e a correlação medida. Um conjunto de valores $X_{inicial,i}$ = 50.000 foram sorteadas a partir de N_i{100,10} e observações $X_{2,i}$ e $X_{1,i}$ são simulados a partir de $X_{k,i} = X_{inicial,i} + N\{0, 5\}$. As variáveis resultantes têm: $\rho\{X_1, X_2\} = 0.62$, $Cov\{X_1,X_2\} = 100$ e $\sigma_{1,real} = \sigma_{2,real} = 11.2$ (**Tabela 3** e **Tabela 4**).

Em seguida, os valores iniciais são usados para simular as observações X_1 e X_2 são perturbadas por erros aditivos ou multiplicativos. O erro total adicional é constante para ambos os casos, sendo $C_k + D = 0.1$ e $A_k + B = 10$. Nós variamos as proporções entre componentes não compartilhados e compartilhados, respectivamente C_k/D e A_k/B em 0, 10, 30, 50, 70, 90 e 100%. A **Tabela 3** e **Tabela 4** mostram os resultados dessas simulações na presença de erro aditivo ou multiplicativo, respectivamente.

Tabela 3 - Covariância e correlação medida com erro aditivo

Estatísticas reais			**Estatísticas observadas - Erro aditivo**			
$Cov_{real}\{X_1,X_2\}$	$\sigma_{1,real} = \sigma_{2,real}$	B	A_k	r^2	$Cov_{obs}\{X_1,X_2\}$	$\sigma_{1,obs} = \sigma_{2,obs}$
100	11.2	0	10	0.19	100	14.2
100	11.2	1	9	0.24	102	14.2
100	11.2	3	7	0.36	109	14.2
100	11.2	5	5	0.52	126	14.2
100	11.2	7	3	0.66	147	14.2
100	11.2	9	1	0.76	180	14.2
100	11.2	10	0	0.79	200	14.2

Covariância e correlação de valores sorteados de $N\{100, 10\}$ e perturbados individualmente por um erro $N\{0, 5\}$, resultando inicialmente em $\rho(X_1, X_2) = 0.62$. Em seguida, esses valores foram perturbados por erros aditivos $A_k + B = 10$, com diferentes proporções de erros compartilhados (B) e não compartilhados (A_k).

Tabela 4 - Covariância e correlação medida com erro multiplicativo

Estatísticas reais			Estatísticas observadas - Erro Multiplicativo			
$Cov_{real}\{X_1,X_2\}$	$\sigma_{1,real} = \sigma_{2,real}$	D	C_k	r^2	$Cov_{obs}\{X_1,X_2\}$	$\sigma_{1,obs} = \sigma_{2,obs}$
100	11.2	0	0.1	0.20	100	14.2
100	11.2	0.01	0.09	0.23	100	14.2
100	11.2	0.03	0.07	0.35	109	14.2
100	11.2	0.05	0.05	0.51	125	14.2
100	11.2	0.07	0.03	0.67	151	14.2
100	11.2	0.09	0.01	0.77	184	14.2
100	11.2	0.1	0	0.79	200	14.2

Covariância e correlação de valores sorteados de *N*{100, 10} e perturbados individualmente por um erro *N*{0, 5}, resultando inicialmente em $\rho(X_1, X_2) = 0.62$. Em seguida, esses valores foram perturbados por erros multiplicativos $C_k + D = 10$, com diferentes proporções de erros compartilhados (D) e não compartilhados (C_k).

Nas Tabelas 3 e 4, a primeira linha é, respectivamente, $A_k = 10$ ou $C_K = 0.1$, ou seja, onde todo o erro é independente do teor e, nesses dois casos, a covariância medida é igual a do fenômeno sem erros, pois a covariância não é afetada por erros independentes do teor. Por outro lado, tanto erros independentes quanto proporcionais ao teor têm o mesmo impacto no desvio-padrão da população e, portanto, quanto maior for A_k ou C_K, menor será a correlação observada (**Tabela 4**).

Já os erros compartilhados têm o impacto inverso: maior for a variância do erro compartilhado, mais próximo $\rho\{X_1, X_2\}$ será de 1, já que a correlação medida combina a correlação real e a correlação compartilhada do erro. Em ambos os casos, negligenciar o tipo de erro associado às observações pode levar a estatísticas não representativas dos fenômenos geológicos de interesse.

4.3 INFERÊNCIA DE PARÂMETROS GEOESTATÍSTICOS MULTIVARIADOS ATRAVÉS DE OBSERVAÇÕES COM ERROS

Realizações do fenômeno real são condicionadas a honrar a estrutura de correlação, o covariograma e o histograma do fenômeno real. A inferência das estatísticas reais a partir de observações com ou sem esses erros leva a resultados diferentes. Esta seção apresenta um fluxo de trabalho de simulação multivariada que lida corretamente com a presença de erros correlacionados entre as variáveis.

4.3.1 Metodologia

O conjunto multivariado $\{z_{k,obs}(\mathbf{u}_i),\ k = 1,...,K,\ i = 1,...,N\}$ é composto por N observações de K variáveis. As observações vem de $z_{k,real}(\mathbf{u}_i) + \varepsilon_k(\mathbf{u}_i)$, onde $\varepsilon_k(\mathbf{u}\ \mathbf{u}_i$ é o erro amostral com variância $\sigma^2_{k,erro}(\mathbf{u})$ que pode ser diferente para cada observação. Assumimos o erro de amostragem sem viés $E\{\varepsilon_k(\mathbf{u}_i)\} = 0$ e independente do teor $Cov\{Z_{k,real}(\mathbf{u}_i), \varepsilon_k(\mathbf{u}_i)\} = 0$. Assim como no método univariado, o fluxo de trabalho simula múltiplas realizações (L) do fenômeno real nas posições amostradas. Em seguida, os bancos de dados simulados são usados para simular o restante do modelo $Z^L_{k,real}(\mathbf{.})$.

i. O erro $\varepsilon_k(\mathbf{u}_i)$ associado a $z_{k,obs}(\mathbf{u}_i)$ é estimado utilizando algum dos fluxos apresentados na Seção 3.2 ESTIMANDO O ERRO ASSOCIADO A CADA DADO. A proporção das diferentes componentes de erro é inferida através do método apresentando no APÊNDICE 2 - ESTIMANDO AS COMPONENTES PROPORCIONAIS E INDEPENDENTES DE ERRO.

ii. O histograma, o covariograma e a estrutura de correlação das variáveis de interesse são inferidos a partir de observações.

iii. As realizações $z^L_{k,real}(\mathbf{u}_i)$ nas posições $\mathbf{u}_i$ são condicionadas a honrar as estatísticas estimadas (ii), os valores previamente simuladas e o intervalo de incerteza centrado em $z^*_{k,real}(\mathbf{u}_i)$ com variância dada por $\varepsilon_k(\mathbf{u}_i)$.

iv. Simular L modelos de $Z_{k,real}(\mathbf{.})$, condicionados às estimativas das estatísticas reais (histograma e covariograma) e aos valores do banco de dados. Como todos os valores utilizados nessa etapa são assumidos como *hard data*, pode-se empregar métodos multivariados de simulação estocástica mais simples, tais como PCA / PPMT (BARNNET; DEUTSCH, 2015) ou SGS com a estimativa das probabilidades locais através de COK. Para cada realização ser simulada, um novo banco de dados deve ser previamente simulado e considerado.

v. Retro-transformação dos valores simulados para as suas unidades originais, pós-processamento e análise das realizações, a fim de avaliar a incerteza e outras estatísticas de interesse.

A implementação deste processo é bastante simples. As seções a seguir fornecem informações detalhadas sobre cada etapa e alguns exemplos.

4.3.2 Estimando o covariograma do fenômeno real

Na presença de erros de correlacionados ao teor, a equivalência entre $C_{real}(\mathbf{h})$ e $C_{obs}(\mathbf{h})$ quando $\mathbf{h} > 0$ é dada pela Equação 22. Quando $\mathbf{h} = 0$, a covariância entre uma amostra e ela mesma é a sua variância (Equação 21). O covariograma do fenômeno real entre duas posições, i e j é inferido através de:

$$C_{real,(i,j)}(\mathbf{h}) = \begin{cases} C_{obs,(i,j)}(\mathbf{h}), & \mathbf{h} > 0 \\ C_{obs,(i,j)}(\mathbf{h}) - B^2 - D^2\{E(Z_{1,obs})E(Z_{2,obs})\}, & \mathbf{h} = 0 \end{cases} \quad (25)$$

Na formulação apresentada da equação 25 é assumida que B e D são constantes no espaço, e, assim, afetam todos os passos do covariograma de forma semelhante, o que possibilita que uma mesma correção seja aplicada para todo o covariograma. No entanto, caso B e D variem no espaço de forma que afetem de forma diferente os passos do covariograma experimental, cada passo deve ser corrigido pela Equação 25 através dos seus valores B e D médios dos pares de amostras utilizadas para o seu cálculo.

A Equação 25 é o modelo geral, o qual os modelos anteriormente apresentados são casos específicos, tais como quando o erro é independente do teor ou quando os erros não são compartilhados. Reorganizando os termos para que possamos inferir o $C_{obs,(i,j)}(\mathbf{h})$ entre dois nós a partir das componentes de erro e do modelo inferido do covariograma do fenômeno real, chegamos a um sistema de krigagem mais geral:

$$\begin{pmatrix} C_{real}(x_1 - x_1)(1 + C^2_1 + D^2) + A^2_1 + B^2 & \cdots & C_{real}(x_1 - x_n) & 1 \\ \vdots & \ddots & \vdots & 1 \\ C_{real}(x_n - x_1) & \cdots & C_{real}(x_n - x_n)(1 + C^2_n + D^2) + A^2_n + B^2 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_n \\ -\mu \end{pmatrix} = \begin{pmatrix} C_{real}(x_1 - x_0) \\ \vdots \\ C_{real}(x_n - x_0) \\ 1 \end{pmatrix} \quad (26)$$

O sistema de krigagem acima (Equação 25) lida com a covariância entre duas observações de duas variáveis diferentes. Nos pares entre observações da mesma variável, é utilizado o sistema previamente apresentado, do caso univariado com erros proporcionais ao teor.

No exemplo apresentado na Seção 3.5.1.3 Exemplo 2 Estimando o variograma real) é ilustrado a capacidade de o método inferir o covariograma cruzado em função dos erros das duas amostras (Equação 25).

4.4 SIMULANDO O FENÔMENO REAL

Tendo desenvolvido um modelo de corregionalização univariado capaz de considerar o impacto do erro amostral associado a cada observação, nós podemos usar esse conhecimento da relação espacial entre qualquer par de nós para definir o teor médio ótimo e a variância mínima em cada posição em que uma observação esteja disponível e, então, essa observação pode ser substituída por um valor simulado a serem assumidos como *hard data*.

As realizações no caso multivariado são condicionadas a reproduzir os modelos inferidos da distribuição, correlação espacial de $Z_{k,real}(\mathbf{.})$ e estrutura de correlação entre as variáveis. Em vez de ser condicionado a honrar $z_{k,obs}(\mathbf{u}_i)$, as realizações consideram o espaço de incerteza $N\{Z^*_{k,real}(\mathbf{u}_i), \sigma^2_{k,krig}(\mathbf{u}_i)\}$, onde a variância de krigagem $\sigma^2_{k,krig}(\mathbf{u})$ é diretamente influenciada pelos valores de $\sigma^2_{k,erro}(\mathbf{u})$ associados ao valor colocalizado e o erro nos vizinhos. Dessa forma, o método é capaz de gerar uma melhor reprodução do fenômeno simulado e avaliar o comportamento do espaço de incerteza entre as realizações e o valor real.

As simulações das bases de dados $z^L_{k,real}(\mathbf{u}_i)$ são um problema de simulação colocalizada, onde observações $z_{k,obs}(\mathbf{u}_i)$ (*soft data*) estão disponíveis em todos os locais a simular $z^L_{k,real}(\mathbf{u})$. A co-krigagem intrínseca colocalizada (ICCK; BABAK; DEUTSCH, 2009a) é considerada no caso multivariado devido à simplicidade e por ser capaz de cossimular funções aleatórias sem exigir a inferência e modelagem de uma matriz completa de covariância. A correlação entre as variáveis de interesse sejam as observações ou *hard data* já simulados são inferidos através da Equação 23 e 24, garantindo a correta correlação entre eles, além disso, o ICCK melhora a reprodução do variograma quando comparado a outros métodos de cossimulação, mesmo quando as diferentes variáveis diferem significativamente na continuidade (BABAK; DEUTSCH, 2009b). A abordagem é composta por duas etapas.

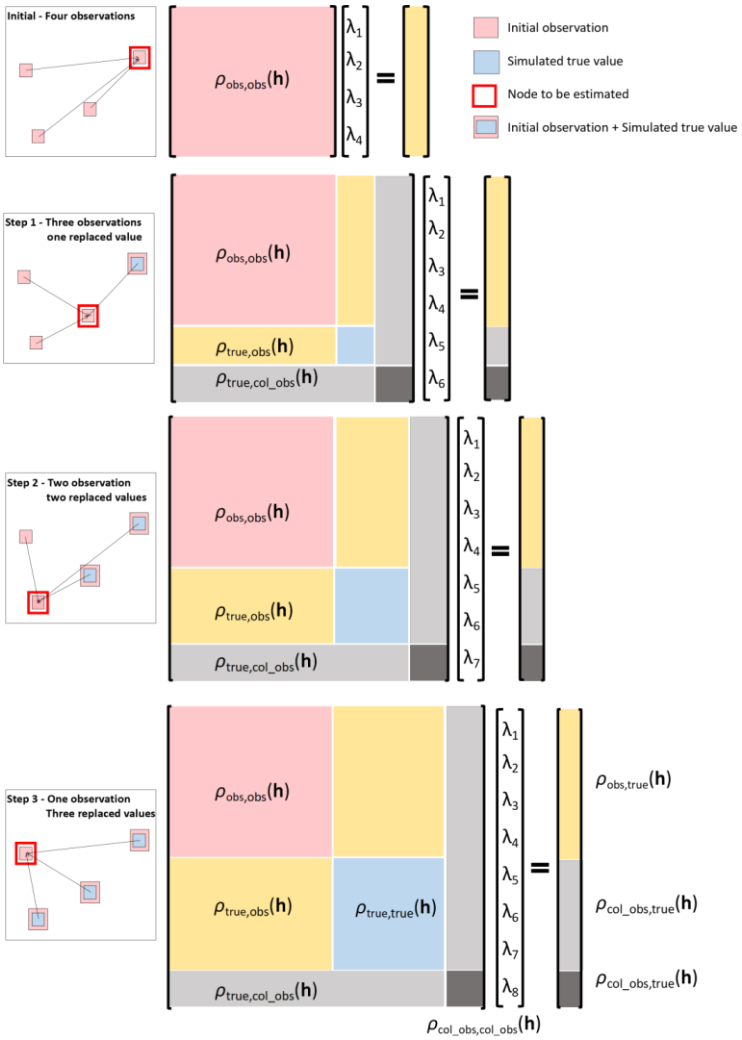
i. O fenômeno real é simulado nas posições amostradas. Uma por uma, cada observação é escolhida aleatoriamente e substituído por um *hard data* simulado. O caminho aleatório deve passar por todos os nós e ser redefinido a cada simulação como forma de evitar artefatos. A probabilidade local é estimada por ICCK usando a observação colocalizado e os *hard data* previamente, além de estar condicionado a

reproduzir a estrutura de correlação inferida entre elas (Figura 33). O covariograma cruzado entre cada par de nós é inferido a partir da Equação 25 em função de seus erros individuais. O processo é repetido para cada variável *K*, sendo as outras combinadas em uma supersecundária (BABAK; DEUTSCH, 2009b).

ii. Após substituir todos valores iniciais por realizações de *hard data*, o conjunto de novos valores é assumido como um novo banco de dados sem erros e é usado para simular todo o modelo. Qualquer algoritmo de simulação condicional multivariado pode ser usado, tais como PCA / PPMT (BARNNET; DEUTSCH, 2015) ou sGs com a estimativa das probabilidades locais através de COK. Cada realização do modelo completo deve ser condicionada por um novo banco de dados.

A Figura 33 mostra quatro etapas na substituição das observações (indicados pela cor rosa) por *hard data* simulados (indicados pela cor azul), considerando também a correlação entre as observações e cada *hard data* simulado (cor cinza).

Figura 33 - Esquema do sistema de ICCK para estimativa das CCDF



Esquema do sistema de ICCK para estimativa das CCDF na simulação do banco de dados. O sistema é composto pelos variogramas diretos e cruzados entre as observações (índice *obs)* e o fenômeno real (índice *true),* assim como entre o valor colocalizado e os *hard data* já simulados. A diagonal (i=j) entre as observações é preenchida pelos erros associados aos dados individuais.

O ICCK pode usar o modelo intrínseco de correlação entre o fenômeno real *K* e suas observações $\rho_{k,Y(obs,real)}$, sendo esta inferida em cada posição **u** em função de erro de amostragem associado às observações colocalizadas através da Equação 23 e 24. A estimativa por ICCK também pode considerar na estimativa observações de um segundo fenômeno *j*, desde que essa ofereça informações adicionais sobre o fenômeno de interesse *K*.

Caso as observações dos fenômenos *K* e *J* tenham seus erros compartilhados, a correlação entre elas pode ser calculada pela Equação 2. Portanto, o erro de estimar uma

variável *K* a partir de uma observação do fenômeno j pode ser obtido através do erro de regressão entre *K* e *J*.

5 CONCLUSÕES E RECOMENDAÇÔES

Esta tese apresentou um minucioso desenvolvimento de um fluxo de trabalho de simulação na ausência do que vem a ser chamado *hard data*, que são medições não afetadas por erros. Nessa condição, as definições do covariograma e da distribuição não são diretas, visto que as observações disponíveis não são medições exatas do fenômeno real a ser simulado, mas sim a combinação deste com uma componente de erro, que pode corresponder a mais de 30% da variância total observada. Neste capítulo, serão apresentadas as conclusões sobre a metodologia, sua aplicabilidade e as contribuições da tese, avaliando as vantagens e limitações de cada metodologia proposta e a sugestões para trabalhos futuros.

5.1 CONCLUSÕES

Não existem *hard data* na mineração. Todas as observações são afetadas por erros. Considerando essa realidade, o objetivo desta tese foi o de propor métodos de simulação capazes de gerar modelos que lidem corretamente com os erros individuais associados a todas as observações disponíveis, levando a possibilidade de se obter realizações que realmente sejam equiprováveis ao fenômeno real e que meçam corretamente a incerteza entre o modelo e o fenômeno real. Em função dos objetivos inicialmente propostos, chegou-se às seguintes conclusões.

- **Definir um modelo do comportamento do erro amostral, que seja válido tanto no caso univariado quanto multivariado**

Foi desenvolvido um modelo que considera a observação conhecida como produto do valor real, que é desconhecido, e uma componente de erros. O modelo foi apresentado em sua versão analítica para os casos univariado e multivariado. O modelo foi apresentado tanto na forma analítica quanto na matricial e assumiu as mesmas hipóteses assumidas na literatura, que são: ausência de viés, distribuição gaussiana do erro e que o erro é composto de uma componente independente e uma dependente. Nesse caso, existe uma relação linear entre erro e a magnitude dos teores.

No entanto, o tema de análise de erros para dados multivariados se mostrou insuficiente para o desenvolvimento do trabalho. Portanto, foi apresentada a possibilidade de que erros sejam compartilhados entre as observações de duas ou mais variáveis. Essa

situação é completamente plausível e ignorar a sua presença pode levar a estatísticas, como covariância, correlação ou variância, completamente diferentes daquela do fenômeno real.

- **Desenvolver alternativas para estimar os parâmetros do modelo de erro**

Após a definição dos modelos de erros, foram apresentadas diferentes alternativas para que os parâmetros fossem estimados. Foram apresentados métodos simples, como inferir o erro médio de um subgrupo de observações através do efeito pepita do variograma ajustado a elas, e o uso da Teoria da Amostragem para inferência do erro, quando o erro fundamental possa ser considerado a única fonte de erros que afete tais dados. Na disponibilidade de amostras duplicatas, foi desenvolvido um método baseado em etapas de análise gráfica e testes de hipótese de forma a definir os subgrupos que minimizem o erro de estimativa.

O uso da variância de krigagem como medida de erro de amostragem foi discutido e descartado, visto o fato de a primeira variar em função da proximidade entre amostras não ser algo plausível para o erro amostral, que é um produto da interação entre o protocolo amostral empregado e o tipo de rocha, mas não da sua posição espacial.

- **Desenvolver alternativas para transformar erros amostrais para unidades Gaussianas**

O uso de métodos geoestatísticos de simulação baseados na hipótese multiGaussiana, torna necessário transformar os valores iniciais para gaussianos. Na presença de erros, é necessário que estes também sejam transformados. A transformação dos erros era um tema sem análise na literatura geoestatística, fazendo necessário inserir a influência do erro nas estimativas de forma indireta na simulação, ou através do emprego de DDS, já que neste caso não se exige dados gaussianos.

Através do desenvolvimento matemático, foi provado que, após a transformação propostas, a proporção da componente de erro em relação à variância global das observações em unidades originais é mantida após a transformação para dados gaussianos. O método é capaz de lidar com todos os tipos de erros aqui analisados, sendo eles transformados para uma única componente independente.

- **Desenvolver formas para estimar a correlação espacial, estrutura de correlação e a distribuição do fenômeno real através de observações com erros**

Na presença de erros de amostragem, as estatísticas ajustadas aos dados são diferentes daquelas do fenômeno real. Condicionar simulações a honrar as estatísticas dos dados leva a realizações que não são equiprováveis ao fenômeno real. Isso posto, os modelos de erros desenvolvidos foram empregados para obter as estatísticas reais através das observações. Foram desenvolvidas as equações para o caso mais geral, que lida com dados multivariados com erros independentes e correlacionados ao teor, sendo eles compartilhados entre variáveis ou não. Os outros dois modelos são casos especiais do primeiro, sendo o caso univariado na presença de erros independentes do teor e o caso de dados univariado correlacionados com o teor.

- **Desenvolver um sistema de krigagem que considere o erro individual de cada observação na definição dos pesos**

Foram desenvolvidos sistemas de krigagem em que o covariograma entre cada par de nós é definido automaticamente em função do covariograma assumido como do processo geológico e dos erros associados as observações. E aqui surge a maior vantagem dos três modelos de corregionalização apresentados: eles podem ser estimados diretamente de $C_{obs,(i,j)}(\mathbf{h})$, desde que os erros associados a tais observações sejam conhecidos. A inclusão dos pesos e erros diretamente no sistema de krigagem garante a minimização do erro de estimativa ao substituir covariância e propriedades médias de um subgrupo pelo erro e estatísticas corretas de cada observação.

- **Desenvolver métodos de simulação estocástica condicionada aos parâmetros do fenômeno real e à incerteza de cada dado.**

Hard data (livres de erro) não podem de fato serem obtidos, mas valores equiprováveis de *hard data* podem ser simulados a partir de observações afetadas por erro de amostragem. Para tal, foi desenvolvido um fluxo de simulação baseado em duas etapas.

Primeiro, são simulados diferentes bancos de dados de *hard data*.Para tal, são empregados os modelos lineares de corregionalização desenvolvidos nessa tese. No caso univariado, é empregado a co-krigagem para definir a CDF local necessária para a simulação. No caso multivariado, é empregada a co-krigagem intrínseca colocalizada devido a sua

simplicidade e por ser capaz de cossimular funções aleatórias sem exigir a inferência e modelagem de uma matriz completa de covariância, tornando o método mais rápido.

Após a simulação de diferentes bancos de dados, cada um é utilizado para uma realização do modelo completo. Nesta etapa, existem apenas *hard data* e, portanto, podemos empregar métodos de simulação estocástica mais simples, tais como sGs com a estimativa das probabilidades locais através de SK. Da mesma forma, o caso multivariado pode empregar métodos de simulação estocástica mais simples, tais como PCA/PPMT ou sGs com a estimativa das probabilidades locais através de CSK.

As realizações do método proposto geram modelos com uma melhor reprodução do fenômeno simulado, assim como um espaço de incerteza mais realista entre as realizações e o valor real.

5.2 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

O desenvolvimento matemático e a apresentação da hipótese desta tese abrem caminho para diversas outras aplicações além das aqui discutidas. São listadas abaixo recomendações para trabalhos futuros.

- **Avaliação do método de inferência do sistema de covariância para o fim de estimativa**

dado o escopo do trabalho, não foram analisados os ganhos dos estimadores para gerarem mapas krigados. No entanto, os sistemas de krigagem desenvolvidos resolvem diversos problemas que haviam entre os métodos geoestatísticos, por exemplo:

i. a restrição de KVME apenas lidar com as observações independentes do teor limita em muitas de suas aplicações, o método proposto que lida com erros independentes e proporcionais ao teor soluciona esse problema;
ii. a maior restrição dos métodos de coestimativa é a necessidade de definir o modelo linear de corregionalização através do ajuste de variogramas diretos e cruzados a todos os subgrupos de dados. Sendo conhecido o erro, o método proposto soluciona o problema de uma forma eficiente, em que apenas um modelo é necessário e o restante dos ajustes são definidos pelo próprio sistema de krigagem. É recomendado que em trabalhos futuros fossem avaliados os ganhos do método proposto em relação aos de coestimativa.

- **Estendendo o método para variáveis categóricas.**

O método foi desenvolvido para variáveis contínuas. No entanto, ele pode ser facilmente estendido às variáveis categóricas através do proposto por Soares *et al*. (2016), em que cada valor é transformado em um indicador *p*(x) que define a probabilidade da observação x pertencer à categoria X. O valor p (x) tem exatamente o mesmo significado de indicador I(x) mas, como a observação tem erros, a codificação pode levar qualquer valor entre 0 e 1 em vez de ser restrito a 0 ou 1. A probabilidade local estimada de uma determinada observação para pertencer a diferentes categorias define o erro local associado a cada informação categórica. Sullivan (1984) e Alabert (1987) apresentaram mais discussões sobre o uso de dados categóricos com erro.

- **Uso do método proposto para otimização de malha de sondagem e seleção do método de amostragem em cada posição.**

A otimização de malha amostral foi um dos temas mais discutidos e publicados na literatura geoestatística nas últimas décadas. Devido à incerteza não ser espacialmente homogênea, a adição de novos furos em áreas de maior incerteza é a forma mais eficiente de reduzir o erro. Entre as metodologias empregadas, a mais comum é o uso de simulação geoestatística para avaliar a incerteza global e local dos valores dos nós do modelo. Assumindo cada realização como um cenário análogo ao real, o nó simulado mais próximo ao local selecionado para receber uma nova amostra é adicionado ao banco de dados inicial e, então, a incerteza de cada bloco é recalculada. A posição pode ser definida simultaneamente para vários furos (KOPPE, 2009; ORTIZ *et al*., 2012) ou furo-a-furo (PILGER *et al*., 2001).

Todos esses métodos consideram apenas a otimização posicional, em que se assume a ausência de erro nas amostras. No entanto, os custos de aquisição de amostras correspondem a uma parcela considerável do orçamento de pesquisa mineral dentro das operações mineiras. Considerando a diversidade de métodos de amostragem, protocolos de preparação e de análise disponíveis, levar em conta a relação entre o erro amostral e o custo de aquisição é uma oportunidade econômica que sempre deve ser investigada (SILVA E COSTA, 2016a). Por exemplo, considerando ser equivalente o custo para realizar 3 furos de sondagem rotativa diamantada, 5 furos de circulação reversa ou 20 canais coletados manualmente, qual desses métodos amostrais proporcionaria mais benefício? E qual seria

a melhor alternativa se pudéssemos reduzir a quantidade de amostras e utilizarmos protocolos amostrais e de preparação diferentes? Essas questões não foram respondidas de forma direta pelos métodos disponíveis para otimização amostral.

Considerando o desenvolvido na presente tese, podemos adicionar essa nova camada de otimização na definição ótima de novas amostras, considerando a existência de alternativas amostrais com diferentes custos de aquisição e erros associados. Isso nos leva a um problema que não é apenas geoestatístico, mas também financeiro entre o custo de aquisição e o retorno esperado de cada nova amostra. Os métodos desenvolvidos na tese possibilitam responder as questões levantadas, sendo recomendado a aplicação em situações de otimização que lide simultaneamente com a escolha da melhor posição para cada amostra a ser coletada, assim como qual o método mais adequado para cada posição. Isso só é possível com um método capaz de considerar a contribuição individual de cada amostra no espaço de incerteza simulado, assim como o de considerar simultaneamente qualquer quantidade de dados com diferentes erros.

- **Possíveis simplificações no fluxo de trabalho.**

O método proposto se mostrou funcional e todas as etapas tiveram suas validades provadas analiticamente, o que torna possível medir o impacto de possíveis simplificações nesse processo.

Uma primeira oportunidade é quando o produto de interesse da simulação são realizações no suporte de blocos. Nesse caso, a incerteza associada a variância nos nós amostrados é minimizada pelo restante dos nós que discretizam o bloco e que não coincidentes com observações. Por exemplo, ao simular um bloco 10x10 x10m discretizado em um grid 1x1 x1m, a influência da pequena fração desses 1000 nós que são coincidentes com observações é diluída no restante dos nós. Portanto, é relevante investigar o emprego de sGs condicionada a honrar apenas os parâmetros globais corretos, mas dispensando a etapa de simulação de banco de dados.

- **Lidando com suportes heterogêneos.**

O método assume um único suporte amostral para todas as observações. No entanto, o fluxo de trabalho e equações apresentadas pode ser adequado para lidar com suportes variáveis, onde o covariograma estimado para erro zero, como desenvolvido ao longo do trabalho, também é levado para a base de suporte unitário. Tal variograma do fenômeno

real em suporte unitário é, então, reescalonado para cada par de nós em função dos seus erros e dos seus suportes.

- **Usar o sistema de krigagem para definir o drift de processos não estacionários.**

Os sistemas de krigagem desenvolvidos podem ser empregados para definir o comportamento regional de processos não estacionários. Para tal, as amostras são assumidas como sendo observações do processo de larga-escala, enquanto as variações de pequena escala (aquelas krigadas em processos estacionários) são considerados como erros de medição. O resultado é a adição de efeito pepita e de variâncias nas diagonais no sistema que leva a um processo mais suavizado e menos sensíveis a pequenas variações, o que é desejado na estimativa do processo de larga escala.

- **Estender as discussões de formas de medir o erro para a área de petróleo.**

A estimativa do erro associado aos dados foi gerada com foco na área de mineração e outros meios cujas suas amostras são obtidas através da coleta de material fragmentado. No entanto, os dados de petróleo são obtidos principalmente através de métodos geofísicos. Portanto, deve ser investigado como estimar os fatores dos modelos de erro apresentados.

REFERÊNCIAS

ABZALOV, M.Z. Quality control of assay data: a review of procedures for measuring and monitoring precision and accuracy. **Exploration Mining Geology**, v. 17, n.3–4, p. 131–144, 2008.

ABZALOV, M.Z. Sampling errors and control of assay data quality in exploration and mining geology. *In:* Ognyan, I., **Applications and Experiences of Quality Control**, London: Intechopen, 2011. cap. 31, p. 611.

ALABERT, F.G. The practice of fast conditional simulations through the LU decomposition of the covariance matrix. **Journal of Mathematical Geology**, v. 19, p. 369-386, 1987.

ARAÚJO, C.P. **Uso de informação secundária imprecisa como distribuição de probabilidade local em conjuntos de dados completamente heterotópicos**. 2019. Tese (Doutorado em Engenharia de Minas, Metalúrgica e de Materiais) – Programa de Pós- Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019.

ARAÚJO, C.P.; COSTA, J.F.C.L. Integration of different-quality data in short-term mining planning. **Rem: Revista Escola De Minas,** v. 6, p. 221-227, 2015.

BABAK, O.; DEUTSCH, C.V. An intrinsic model of coregionalization that solves variance inflation in collocated cokriging. **Computers; Geosciences**, v.35, p. 603-614, 2009a.

BABAK, O.; DEUTSCH, C.V. Collocated cokriging based on merged secondary attributes. **Mathematical Geosciences,** v. 41, n.8, p. 921-926, 2009b.

BARNETT, R.; DEUTSCH, C.V. Multivariate imputation of unequally sampled geological variables. **Mathematical Geosciences**, v. 47, n.7, p. 791-817, 2015.

CORNAH, A.; MACHAKA, E. Integration of imprecise and biased data into mineral resource estimates. **Journal of the Southern African Institute of Mining and Metallurgy**, v.115, n.6, p. 523-530, 2015.

CRESSIE, N. **Statistics for spatial data**. ed. rev. New York: Wiley-Interscience, 1993. 900 p.

CUBA, M.; LEUANGTHONG, O.; ORTIZ, J.M. Accounting for different sampling errors in exploratory drilling campaigns in resource/reserves modelling. *In*: CCG Annual Meeting, 10, 2008, Edmonton, **CCG Annual Report 10**, paper 306, 2008.

CUBA, M.; LEUANGTHONG, O.; ORTIZ, J.M. Transferring sampling errors into geostatistical modelling. **Journal of the Southern African Institute of Mining and Metallurgy**, v. 112, n. 11, p. 971-983, 2012.

DELHOMME, J.P. **Application de la théorie des variables régionalisé es dans les sciences de l'eau.** 1976. Tese (Doutorado Universidade de Paris VI) – Fountainebleau, 1976.

DERAISME, J. (2009) Estimation of iron ore resources integrating diamond and percussion drillholes. *In*: 34th **International Symposium on Computer Applications in the Mineral Industries (APCOM2009)**, Vancouver. Proceedings, 2009, p. 1-8.

EMERY, X. Geology resource and reserve evaluation in the presence of imprecise data. **CIM Bulletim**, v. 90, n.1089, p. 366–377, 2005.

FRANCOIS-BONGARCON, D. Error variance information from paired data: applications to sampling theory. **Exploration and mining Geology**, v. 7, p. 161–165, 1998.

FRANCOIS-BONGARCON, D. Theory of sampling and geostatistics: an intimate link. **Chemometrics and intelligent laboratory systems.** v.74, n.1, p.143–148, 2004.

FROIDEVAUX, R. Probability field simulation. *In*: Geostatistics Troia '92, 1993, Troia, **Proceedings**, Dordrecht : Springer, 1993, v.1, p.73–84.

GOOVAERTS, P. **Geostatistics for natural resources evaluation**. Applied Geostatistics Series. New York: Oxford University Press, 1997.

GOOVAERTS, P. Ordinary CoKriging revisited. **Mathematical Geology**, v.30, n.1, p.21-42, 1998.

GY, P. **Sampling of particulate materials—Theory and Practice**. 2. ed. Amsterdan: Elsevier, 1982.

HOWARTH R.; THOMPSON M. Duplicate analysis in geochemical practice, part 2: examination of vario proposed method and examples of its use. **Analyst**, v.101, p.699–709, 1976.

ISAAKS, E.H. **The application of monte-carlo methods to the analysis of spatially correlated data.** 1990. Tese (Doutorado). Universidade de Stanford, 1990.

ISAAKS, E.H.; SRIVASTAVA, R.M. **An introduction to applied geostatistics**. Oxford: Oxford University Press, 1989. 267 p.

JOURNEL, A.G. **Simulations conditionnelles--throrie et pratique.** 1974. Tese (Doutorado Universidade de Paris VI) – Fountainebleau, 1974.

JOURNEL, A.G. Constrained interpolation and qualitative information: the soft kriging. **Mathematical Geology,** v. 18, p.269-286, 1986.

JOURNEL, A.G. Probability fields: another look and a proof. *In*: Stanford Center For Reservoir Forecasting, **Report 8**. Stanford, 1995.

JOURNEL, A.G.; HUIJBREGTS, C. **Mining geostatistics**. New York: Academic Press, 1978.

KOPPE, V. **Metodologia para Comparar a Eficiência de Alternativas para Disposição de Amostras.** 2009. Tese (Doutorado em Engenharia de Minas, Metalúrgica e de Materiais) – Programa de Pós- Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

KRIGE, D.G. A Statistical Approaches to Some Basic Mine Valuation Problems on the Witwatersrand. **Journal of the Chemical, Metallurgical and Mining Society of South Africa,** v. 52, p.119-139, 1951.

MAGRI, E.; ORTIZ, J. (2000) **Estimation of economic losses due to poor blast hole sampling in open pits**. *In:* Geostatistics, 2000, Proceedings of the 6th International Geostatistics Congress, 2000, Cape Town, 2000. v.2, p.732–741.

MARCOTTE, D. Conditional simulation with data subject to measurement error: Post-simulation filtering with modified factorial kriging. **Mathematical Geology,** v.27, p.749–762, 1995.

MARECHAL, A. **Cokrigeage et regression en correlation intrinsique.** 1970. Tese (Doutorado). Universidade de Paris VI, Fountainebleau, 1970.

MATHERON, G. Principles of Geostatistics. **Economic Geology,** v. 58, p.1246-1266, 1963.

MATHERON, G. Forecasting block concentration distributions: the transfer functions. **Advanced Geostatistics in the Mining Industry.** Dordrecht:Reidel, p.237-251, 1976.

MONTGOMERY, D.C. **Introduction to statistical quality control**. New York: Wiley, 2009.

MONTGOMERY, D.C.; RUNGER, G.C. **Applied statistics and probability for engineers**. York: Wiley, 2014.

ORTIZ, J.M.; MAGRI, E.J.; LÍBANO, R. Improving financial returns from mining through geostatistical simulation and the optimized advance drilling grid at el tesoro copper mine. **Journal Of The Southern African Institute Of Mining And Metallurgy**, v.112, n.1, p.15-22, 2012.

PILGER, G.G.; COSTA J.F.C.L.; KOPPE J.C. Additional samples: where they should be located. **Natural Resources,** v.10, p.197–207, 2001.

PITARD, F.F. **Pierre Gy's sampling theory and sampling practice. Heterogeneity, sampling correctness and Statistical Process Control**. Boca Raton: CRC Press, 1993.

PYRCZ, M.J.; DEUTSCH, C.V. Two artifacts of probability field simulation. **Mathematical Geology**, v.33, n.7, p. 775–799, 2001.

RAMSEY, M.H.; ELLISON, S.L.R. **Eurachem/EUROLAB/CITAC/Nordtest/AMC Guide: Measurement uncertainty arising from sampling: A guide to methods and approaches**, 2007. Disponível em: <http://www.eurachem.rog/guides/ufs.2007.pdf > Acesso em: 18 abr. 2017.

REN, W. **Exact downscaling in reservoir modeling.** 2007. Tese (Doutorado). Universidade de Alberta, Edmonton, 2007.

SAITO, H.; GOOVAERTS P. Accounting for measurement error in uncertainty modeling and decision-making using indicator kriging and p-field simulation: application to a dioxin contaminated site. **Environmetrics**, v.13, n.5–6, p.555–567, 2002.

SANDJIVY, L. **The factorial kriging analysis of regionalized data: Its application to geochemical prospecting.** Geostatistics for Natural Resources Characterization. Part 1, Dordrecht:Reidel, 1984. p. 559–571.

SILVA, V.M. **Análise de sensibilidade das estimativas ao erro amostral, posicional e suas aplicações.** 2015. Dissertação (Mestrado em Engenharia de Minas, Metalúrgica e de Materiais) – Programa de Pós-Graduação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2015.

SILVA, V.M.; COSTA, J.F.C.L. Selecting the maximum acceptable error in data minimising financial losses. **Applied Earth Science**, v.125, n.4, p.214-220, 2016a.

SILVA, V.M.; COSTA, J.F.C.L. Sensitivity analysis of ordinary kriging to sampling and positional errors and applications in quality control. **REM, International Engineering Journal,** v.69, n.4, p. 491-496, 2016b.

SILVA, V.M.; COSTA, J.F.C.L. Using QC data to estimate the individual precision of original samples: a duplicates-based approach. **Applied Earth Science**, v.127, n.3, p.106-112, 2018.

Victor Miguel Silva & João Felipe Coimbra Costa Leite (2020) Sampling error correlated among observations: origin, impacts, and solutions, Applied Earth Science, DOI: 10.1080/25726838.2020.1727126

SOARES, A. Direct sequential simulation and cosimulation. **Mathematical Geology,** v. 33, n.8, p.911–926, 2001.

SOARES, A. et al. Integration of uncertain data in geostatistical modelling. **Mathematical Geosciences**, v.49, n.2, p.253–273, 2016.

SRIVASTAVA, R.M. **Reservoir characterization with probability field simulation.** *In*: SPE Annual Technical Conference and Exhibition, 1992, Washington p. 927–938.

STANLEY, C.R. On the special application of Thompson-Howarth error analysis to geochemical variables exhibiting a nugget effect. **Geochemical Exploration Environmental Analysis**, v.6, p.357–368, 2006.

STANLEY, C.R.; LAWIE D. Thompson-Howarth error analysis: unbiased alternatives to the large-sample method for assessing non-normally distributed measurement error in geochemical samples**. Geochemical Exploration Environmental Analysis**, v.7, p. 1–10, 2008.

SULLIVAN, J. Conditional recovery estimation through probability kriging—theory and practice. **Geostatistics for Natural Resources Characterization**: Dordrecht:Reidel, v.1, p. 365–384, 1984

THOMPSON, M.; HOWARTH, R. The rapid estimation and control of precision by duplicate determinations. **Analyst**. v.98, p.153–160, 1973.

THOMPSON, M.; HOWARTH, R. Duplicate analysis in geochemical practice, part 1: theoretical approach and estimation of analytical reproducibility. **Analyst.** v.101, p.690–698, 1976.

THOMPSON, M.; HOWARTH, R. A new approach to the estimation of analytical precision. **Journal of Geochemical Exploration**. v.9, p.23–30, 1978.

WACKERNAGEL, H. **Multivariate Geostatistics: An Introduction with Applications.** Berlin: Springer, 2003.

ZAGAYEVSKIY, Y.V.; DEUTSCH, C.V. A short note on an update to the change of support program. *In*: CCG Annual Meeting, 13, 2011, Edmonton, **CCG Annual Report 13**, paper 313, 2011.

ZHU, H.; JOURNEL, A. **Formatting and integrating soft data: stochastic imaging via markov–bayes algorithm.** *In:* Geostatistics Troia ’92, 1993, Troia, Proceedings, Dordrecht: Springer, 1993, v.1, p.1–12.

APÊNDICE 1 – ESTIMANDO A CDF ATRAVÉS DE DADOS COM VIÉS

O trabalho é desenvolvido sob a condição de $E\{\varepsilon_{obs}(\mathbf{.})\} = 0$, ou seja, que os dados não são enviesados. Tal decisão é considerada, principalmente, na definição do método de krigagem empregado para estimar a probabilidade local de cada nó a ser simulado através dos *soft data* disponíveis e os *hard data* simulados. Assumindo $E\{\varepsilon_{obs}(\mathbf{.})\} = 0$, a CoK condiciona a soma dos pesos atribuídos aos *soft data* e *hard data* a ser 1. No entanto, essa condição deve ser alterada na presença de dados com viés, abaixo são discutidas alterações no CoK necessários quando o viés é conhecido e quando não o é.

Quando $E\{\varepsilon_{obs}(\mathbf{.})\}$ é conhecido, toda a população é corrigida antes de ser transformada em unidades gaussianas. Sendo $N\{E\{\varepsilon_{obs}(\mathbf{.})\}; \sigma^2_{obs}(\mathbf{.})\}$ temos $E\{\varepsilon_{obs}(\mathbf{.})\} = E\{z_{obs}(\mathbf{.}) - Z_{real}(\mathbf{.})\} = Média_{obs} - Média_{real}$ e podemos obter uma estimativa da distribuição sem viés através de $N^*\{E\{\varepsilon_{obs}(\mathbf{.})\} - Média_{obs}; \sigma^2_{obs}(\mathbf{.})\}$. Todas as outras etapas são empregadas da forma proposta ao longo da tese.

O segundo caso é que quando $E\{\varepsilon_{obs}(\mathbf{u})\}$ é desconhecido, a influência do bias é anulado através da restrição de que a soma dos pesos atribuídos aos *soft data* sejam 0. Feito isso, o algoritmo proposto na metodologia deve ter seu passo 2 adaptado.

Nessa etapa, as observações são substituídas pelas realizações de possíveis valores livres de erros $z^L_{real}(\mathbf{u})$. No entanto, na substituição de um primeiro valor em uma dada vizinhança, só é possível empregar CoK com a condição da soma dos pesos dos *hard data* a 1 quando há ao menos um valor desse tipo. A geração do primeiro valor de hards-data, portanto, pode ser obtido escolhendo uma observação $z_{obs}(\mathbf{u})$ nesta vizinhança e então, substituído por um hard-data sorteado de $N\{z_{obs}(\mathbf{u}), \sigma^2_{erro}(\mathbf{u})\}$ em vez de ser obtido por CoK.

APÊNDICE 2 - ESTIMANDO AS COMPONENTES PROPORCIONAIS E INDEPENDENTES DE ERRO

A componente combinada $B^2 + A_k^2$ controla toda a variância quando as observações medem uma amostra sem o elemento de interesse. Podemos encontrar o valor de $A_1^2 - A_2^2$ computando a variância da diferença entre dois conjuntos de teor nulo $\text{Var}\{z_{1,obs}(\mathbf{0})\} - \text{Var}\{z_{2,obs}(\mathbf{0})\}$ (Equação 1). A Figura 2.1 ilustra a solução.

$$\text{Var}\{z_{1,obs}(\mathbf{0})\} - \text{Var}\{z_{2,obs}(\mathbf{0})\} = A_1^2 + B^2 - A_2^2 - B^2 = A_1^2 - A_2^2 \quad (1$$

À primeira vista, o melhor método para definir a variância quando o teor é nulo é medir a variância das amostras, geralmente, chamadas de “brancas" (sem teor). Comumente, rotinas QC na indústria de mineração usam materiais sem o teor de interesse (por exemplo, quartzo) em vez de amostras do mesmo material e matriz das rochas analisadas. Brancos confeccionados de outros materiais são mais baratos e adequados para verificar a contaminação laboratorial e verificar o manuseio correto das amostras. No entanto, a matriz desempenha um papel importante no comportamento de erro e da sua variância e, portanto, a variância de brancos com matriz diferentes é inadequada na abordagem proposta.

As equações de FSE não podem inferir o componente de erro absoluto. O FSE é nulo quando uma determinada amostra só tem estéril/ganga. Sugerimos modelar a linha de regressão entre a observação e variância de observações de baixo teor e, em seguida, extrapolar a linha para a interseção com o eixo y.

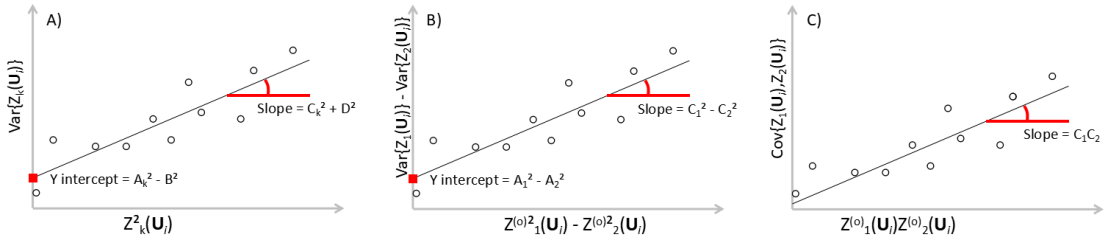
Estimando os componentes C_k e D: pares de duplicatas precisam ser gerados utilizando os mesmos procedimentos utilizados nas amostras originais. As boas práticas sugerem a coleta de duplicatas em todas as etapas de redução e divisão mássica na fase de preparação, permitindo monitorar a variância adicionada por cada etapa. A duplicata de uma amostra pulverizada só será afetada pelas etapas subsequentes do processo. Furos gêmeos, dois canais amostrados ao mesmo tempo e duas partes de um mesmo testemunho são os melhores dados para medir a contribuição combinada de todas as fontes de erros.

A inclinação da regressão linear não é afetada pelas componentes independentes do teor, que são A_k e B. A inclinação pode ser usada para inferir a relação entre erro e teor das componentes dependentes do teor (C_k e D). Três regressões diferentes nos permitem

estimar os valores de $C_1^2 + D^2$, $C^2_2 + D^2$, $C^2_1 - C^2_2$ e C_1C_2. Os valores individuais são definidos resolvendo um sistema de equações lineares.

- A linha de regressão entre a variância e seu valor médio de cada par combinado tem uma inclinação de $C_k^2 + D^2$(Figure 2.1A).
- A regressão entre a diferença $Var\{z_{1,obs}(u_j)\} - Var\{z_{2,obs}(u_j)\}$ (computado com duplicatas de cada variável) e a diferença do teor das amostras originais têm uma inclinação de $C_1^2 - C_2^2$ (Figure 2.1B).
- A regressão entre a covariância das medidas e o produto do teor de amostras originais (ou duplicadas) de duas variáveis observadas no $\mathbf{u}_i$. A inclinação de regressão é C_1C_2 (Figura 2.1C).

Figura 2.1 – Linha de regressão



Linha de regressão e equação de inclinação entre variância e dados são apresentados em três configurações diferentes. A) O eixo Y é a variância de cada par combinado (duplicado e original) e do eixo X sua média; B) Y-eixo é a diferença da variância $Var\{z_{1,obs}(u_i)\} - Var\{z_{2,obs}(u_j)\}$ medido pelos pares combinados de cada variável e X-eixo a diferença da teor de observações originais; C) O eixo Y é covariância entre observações de duas variáveis e eixo X o produto da teor de suas observações originais.

ANEXO I –

Selecting the maximum acceptable error in data minimising financial losses

DISPONÍVEL EM:
HTTPS://WWW.TANDFONLINE.COM/DOI/ABS/10.1080/03717453.2016.1230972?JOURNALCODE=YAES20

ANEXO II –

Using QC data to estimate the individual precision of original samples: a duplicates-based approach

Disponível em:

HTTPS://WWW.TANDFONLINE.COM/DOI/ABS/10.1080/25726838.2018.1497245

ANEXO III –

Sampling error correlated among observations: origin, impacts, and solutions

Disponível em:

https://www.tandfonline.com/doi/abs/10.1080/25726838.2020.1727126