UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE MINAS, METALÚRGICA E DE MATERIAIS (PPGE3M)

SIMONE GUIMARÃES PEREIRA

Inserção de dados faltantes não aleatórios para estimativa de variável geometalúrgica

Porto Alegre

2019

SIMONE GUIMARÃES PEREIRA

GEÓLOGA

Inserção de dados faltantes não aleatórios para estimativa de variável geometalúrgica

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Minas da Universidade Federal do Rio Grande do Sul, como requisito parcial à obtenção do título de Mestre em Engenharia, modalidade Acadêmica.

Orientador: Prof. Dr. João Felipe Coimbra de Leite Costa

Porto Alegre

2019

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

Orientador:

Prof. Dr. João Felipe Coimbra Leite Costa

Banca Examinadora:

Prof. Dr. Rodrigo de Lemos Peroni

Dra. Camilla Zacché da Silva

Dr. Áttila Leães Rodrigues

Prof. Dr. Carlos Pérez Bergmann
Coordenador do PPGEM

DEDICATÓRIA

À minha família

AGRADECIMENTOS

Ao Prof. Dr. João Felipe Costa pela orientação e apoio.

A todos que de alguma forma ajudaram na elaboração desse trabalho, em especial a Drª. Camilla Zacche.

Aos colegas Ronald e Reinaldo pelas horas de discussões.

À Mosaic fertilizantes pela concessão dos dados e colegas de trabalho pelo suporte e apoio.

À minha família e amigos pela paciência.

EPÍGRAFE

“Eu quase que nada não sei. Mas desconfio de muita coisa.”

Guimarães Rosa

RESUMO

Para um bom aproveitamento dos recursos minerais, é necessário potencializar o controle geológico e metalúrgico vinculado a um bom planejamento de lavra. Assim, é possível uma previsão acurada de produção da usina de beneficiamento.

Para isso, é imprescindível um modelamento robusto do depósito. Esse modelo deve contemplar as variáveis químicas (variável preditora) e a relação dessas variáveis com a recuperação mássica do recurso (variável resposta). Por depender da relação com outras variáveis a recuperação mássica é não aditiva. Essa não aditividade deve ser respeitada utilizando metodologias adequadas para modelá-la.

Outro fator importante é a relação numérica entre dados metalúrgicos e dados químicos. Geralmente, a coleta e análise de dados metalúrgicos é inexistente, ou significativamente menor que o número de amostras químicas, e geralmente sua disposição espacial é concentrada apenas nas regiões de alto teor. O que caracteriza os dados como faltantes não aleatórios (MNAR). Isso pode dificultar ou impedir a integração bem-sucedida da variável metalúrgica ao modelo através de métodos geoestatísticos.

A primeira etapa deste trabalho foi complementar os dados faltantes. O método utilizado para inserção é a atualização bayesiana, com transformação fixa dos resultados MAR (mecanismo de falta aleatória) em MNAR (mecanismo de falta não aleatória). Após o banco completo, a estimativa da recuperação respeitou sua não aditividade, utilizando a metodologia de estimada por regressão de Esperança Condicional Alternada (Alternating Conditional Expectation - ACE).

A partir dos métodos listados, foi criado um modelo geometalúrgico. Com o modelo geometalúrgico, foram construídas pilhas de homogeneização e a partir dos valores obtidos na usina de beneficiamento foi feito uma reconciliação. Adicional a essa validação, foram aplicadas as metodologias de estimativa às novas amostras de laboratório. Assim, pôde-se comparar os valores estimados com o modelo de recuperação mássica, contra os valores obtidos com o teste de laboratório.

A reconciliação das pilhas processadas na usina demonstrou que a inserção de dados melhorou a acuracidade e precisão das estimativas do modelo geometalúrgico, obtendo correlação de 0,73 e erro relativo de 1,65, comparado ao modelo gerado com o banco original com dados faltantes não aleatórios, que obteve correlação de 0,65 e erro relativo de 5,85.

Palavras-chave: Inserção, geometalurgia, recuperação, regressão.

ABSTRACT

For mineral resource, it is necessary to strengthen geological and metallurgical control linked to good mining planning. Thus, an accurate production forecast of the beneficiation plant is possible.

For this, a robust deposit modeling is essential. This model should consider the chemical variables (predictor variable) and the relation of these variables to the mass recovery of the resource (response variable). Because it depends on the relationship with other variables, the mass recovery is non-additive. This non-additivity must be respected using appropriate methodologies to model it.

Another important factor is the numerical relationship between metallurgical data and chemical data. Generally, the sampling and analysis of metallurgical data is non-existent, or significantly less than the number of chemical samples, and generally their spatial arrangement is concentrated only in high grade zones. This characterizes the data as non-randomly missing (MNAR). This may hinder the successful integration of the metallurgical variable into the model through geostatistical methods.

The first step of this work was to complement the missing data. The method used for imputation is the Bayesian update, with fixed transformation of the MAR results (missing at random mechanism) into MNAR (missing not at random mechanism). After the dataset is complemented, the estimate of recovery mass performed using the Alternating Conditional Expectation (ACE) regression, respected its non-additivity.

With the above mentioned methods, a geometalurgical model was built. From this model, homogenization piles were generated, and from the values obtained in the processing plant a reconciliation was made. In addition to this validation, the estimation methodologies were applied to the new laboratory samples. Thus, it was possible to compare the values estimated with the mass recovery model, against the values obtained with the laboratory test.

The reconciliation of the processed piles at the plant showed that the data imputation improved the accuracy and precision of the estimates of the geometallurgical, with 0.73 of correlation and 1.65 of relative error compared to the model generated with the original dataset with missing data not at random, with 0.65 of correlation and 5.85 of relative error.

key-words: Imputation, geometallurgy, recovery, regression.

SUMÁRIO

CAPÍTULO 1 APRESENTAÇÃO DO TRABALHO

1.1 INTRODUÇÃO

Na mineração, a maximização do valor econômico do empreendimento é determinante para sobrevivência do mesmo. Para tal, é necessário um melhor aproveitamento dos recursos minerais, que é potencializado quando há controle geológico e metalúrgico vinculados a um bom planejamento de lavra. Assim, é possível uma previsão acurada de produção da usina de beneficiamento.

O estudo de múltiplas variáveis de um depósito e a relação entre essas podem ter um grande impacto na previsibilidade de um empreendimento mineral. Por exemplo, a distribuição espacial e relação do recurso mineral e das variáveis contaminantes, ditam sua possível recuperação na usina de beneficiamento. Para prever com precisão a recuperação do recurso mineral na usina, é imprescindível um modelo robusto do depósito. Esse modelo deve contemplar, além das variáveis contaminantes e do teor do bem mineral, a relação dessas variáveis com a recuperação mássica e metalúrgica do recurso.

A recuperação mássica (RMTOT) é uma variável metalúrgica de interesse nesse estudo, a qual representa a eficiência do processo de concentração é defina como a massa de produto gerado na usina (concentrado) dividida pela massa de alimentação da usina (ROM). E a sua relação com as variáveis químicas de teor, tanto do mineral minério, quanto dos contaminantes, é multivariada. Assim, um modelo de depósito, que contemple a relação multivariada entre variáveis químicas e metalúrgicas é definido como modelo geometalúrgico.

A geração de um modelo geometalúrgico é complexa, pois se baseia na relação multivariada não linear das variáveis químicas (variáveis preditoras) com a variável metalúrgica (variável resposta). Ainda, a variável recuperação mássica (resposta) é não aditiva, pois o seu valor depende da relação com outras variáveis. Essa não aditividade deve ser respeitada utilizando metodologias adequadas para modelá-las.

Porém, há um ponto crucial sobre a complexidade do modelo geometalúrgico, que na maioria das vezes é desconhecido e ignorado: a relação numérica entre dados metalúrgicos e dados químicos. Na maioria dos casos, a coleta e análise de dados metalúrgicos é inexistente, ou quando ocorre, é significativamente menor que o número de amostras químicas, e geralmente sua disposição espacial é concentrada apenas nas

regiões de alto teor. Isso pode dificultar ou impedir a integração bem-sucedida da variável metalúrgica ao modelo através de métodos geoestatísticos.

Dessa forma, um banco de dados com diferenças numéricas entre os dados metalúrgicos e os dados químicos pode ser caracterizado como banco de dados heterotópico. No caso em questão, essa diferença é marcada pela não coleta de dados metalúrgicos nas regiões onde se espera que os valores dos mesmos sejam baixos, pois a caracterização em planta piloto é um processo caro e que demanda tempo. Então a não caracterização dessas amostras ocorre como medida de economia. Assim quando o valor faltante ocorre em virtude do próprio valor amostral, ou seja, a ausência de resultado metalúrgico nas amostras de baixo teor do mineral minério, pode-se dizer que os dados metalúrgicos são faltantes não aleatórios.

A utilização de um banco com dados faltantes não aleatórios, na geração de um modelo geometalúrgico através de métodos geoestatísticos, gera resultados inacurados. Portanto essa falta deve ser tratada inicialmente. Uma forma simples para remediar o problema seria eliminar todas as amostras onde houvesse falta da variável metalúrgica. Apesar de tornar o banco de dados isotópico, a eliminação de amostra é um desperdício de dinheiro já gasto, além disso, manteria o viés amostral, pois continuaria com amostras somente em certas classes de valores, tornando os resultados da estimativa geoestatística ainda mais inacurados.

Outra maneira de tratar dados faltantes é a inserção, a qual completa os dados faltantes tornando o banco de dados isotópico. Existem várias técnicas de complementação amostral disponíveis adequadas para os diferentes tipos de mecanismo de falta que serão detalhados no item 1.2. Barnett e Deutsch (2014) citaram que Enders (2010) apresentou, de forma sucinta, os métodos mais comuns, por ordem de complexidade: (i) média aritmética, que substitui os valores faltantes pela média global da variável, (ii) regressão, que substitui valores faltantes com base em um modelo de regressão e nos valores amostrados, (iii) regressão estocástica é o mesmo que a regressão acima, porém aplica-se amostragem estocástica para adicionar variabilidade realista a cada valor previsto, (iv) imputação múltipla (MI), que infere um modelo dos valores faltantes por amostragem estocástica gerando múltiplas realizações dos dados completos, onde a análise estatística padrão pode ocorrer nos dados completos, antes de combinar os resultados para formar uma estimativa, (v) estimativa de máxima verossimilhança (MLE), que estima uma população de parâmetros não conhecidos para maximizar a log-probabilidade de cada observação. Na análise dos dados faltantes, esses parâmetros são

estimados através de otimizações interativas usando vários subconjuntos dos dados. E (vi) Atualização Bayesiana (AB), que é baseada em MI, porém considera a correlação espacial entre a variável faltante e uma série de variáveis secundárias correlacionadas entre si. Os métodos de (i) a (iii) são adequados para mecanismo de falta completamente aleatório (MCAR) e os (iv) a (vi) para mecanismo de falta aleatório (MAR). Quando o mecanismo de falta é não aleatório (MNAR), os métodos listados acima não são adequados para serem utilizados de forma livre. Para isso, Rubin (1987) propôs uma aplicação simples: quando o mecanismo de falta atuante no conjunto amostral for MNAR, criam-se valores complementados por meio de métodos desenvolvidos para mecanismo MAR. E após, aplicar-se-á uma transformação fixa aos valores complementados a fim de transformá-los em saídas MNAR. Neste trabalho o método utilizado para imputação dos dados faltantes é a Atualização Bayesiana (Ren, 2007), com transformação fixa dos resultados MAR em MNAR, que serão detalhados no capítulo 2.

Tendo um banco de dados, onde a variável metalúrgica em questão está amostrada como as demais, é possível partir para etapa de transposição desta variável da escala de laboratório (amostra) para escada de produção (modelo geometalúrgico). Nessa nova etapa surge o problema da não aditividade dessa variável.

A não aditividade da variável metalúrgica deve ser respeitada utilizando métodos geostatísticos para a estimativa que carreguem o comportamento não linear minimizando possíveis vieses. As metodologias mais adequadas para isso são: regressões, simulação, krigagem de indicadores dentre outros métodos. Neste trabalho, a variável metalúrgica, recuperação mássica da apatita, será estimada por regressão de Esperança Condicional Alternada (*Alternating Conditional Expectations* - ACE) (Breiman & Friedman, 1985). Trata-se de uma técnica não paramétrica que transforma a variável resposta (recuperação mássica) e as variáveis preditoras (teores químicos) para maximizar a correlação da regressão entre elas.

Os métodos listados serão aplicados para gerar um modelo geometalúrgico, que será utilizado para prever os valores de recuperação em pilhas de homogeneização. Os valores das recuperações obtidas serão comparados com os resultados da usina de beneficiamento. Além disso, para validar a não influência de possíveis descontinuidades operacionais da usina, os métodos de estimativa também serão aplicados às novas amostras de laboratório que não foram consideradas no estudo. Assim pode-se comparar os valores estimados com o modelo de recuperação mássica, contra os valores obtidos com o teste de laboratório.

1.2 ESTADO DA ARTE

Um problema comum vinculado à modelagem geometalúrgica é a heterotopia analítica e até mesmo amostral entre as variáveis em estudo. Geralmente, variáveis metalúrgicas são medidas em um suporte muito maior do que as variáveis geológicas, o que pode resultar em um número maior das amostras geológicas do que metalúrgicas (Boisvert *et al*., 2013; Deutsch, 2013; Rossi e Deutsch, 2014 *in* Vieira, 2016). A ignorância dessa incompletude de variáveis na estimativa por métodos geoestatísticos, principalmente quando essa falta é não aleatória, gera viés no modelo. Assim, torna-se necessária a transformação do banco de dados heterotópico em isotópico, que pode ser feito de duas formas: descartando as amostras incompletas ou realizando a complementação por inserção das faltantes. A primeira pode acarretar em construção de um modelo com viés devido à diminuição da quantidade de informação. A segunda deve ser aplicada respeitando a correlação espacial, variabilidade e estatística dos dados, para então obter um modelo que represente o depósito.

Metodologias que realizem a completude de dados vêm sendo estudadas e desenvolvidas ao longo do tempo. Rubin (1976) apresentou a teoria da imputação de dados e relacionou os tipos de relação de falta. Determinando que o conhecimento do tipo de falta de um banco de dados é imprescindível para a escolha adequada de técnicas de inserção.

Rubin (1976) classificou os modelos de relação entre dados faltantes e variáveis medidas em três tipos: os faltantes completamente aleatórios (*Missing Completely at Random* – MCAR), os faltantes aleatórios (*Missing at Random* – MAR) e faltantes não aleatórios (*Missing Not at Random* – MNAR). Para cada caso de falta de dados há uma maneira de tratá-la, de modo que a inserção de dados, para completa-los, não gere viés significativo às estimativas finais.

Quando a relação de falta é completamente aleatória (MCAR) metodologias de inserção clássicas como: média aritmética, regressão linear e regressão estocástica podem ser utilizadas sem a gerar viés significativo ao modelo resultante. Porém, para tratamento de dados faltantes com tipo de relação da falta aleatória MAR essas metodologias não são recomendadas.

A Estimativa por máxima verossimilhança (MLE) é uma metodologia que pode ser aplicada a mecanismo de falta MCAR e MAR, pois realiza estimativa completando

os dados sem geração de viés. A metodologia parte da ideia de especificar a distribuição que melhor ajusta à população das amostras. Assim, tem como objetivo a estimativa dos parâmetros da distribuição dos dados e uma vez que estes são determinados para o melhor ajuste da distribuição, as amostras faltantes são preenchidas por simulação estocástica.

Já Rubin (1978) aborda a Inserção Múltipla (*Multiple Imputation* – MI), para tratamento de dados faltantes MAR, a qual assume que os dados têm distribuição normal. A realização da MI baseia-se na criação de diversos bancos de dados únicos e completos e pode ser dividida em três etapas: a primeira da inserção, a segunda da análise e a terceira da combinação. A etapa da inserção é subdividida em dois passos (I e P).

I: é construído um conjunto de regressões para formar um banco de dados completo, com os faltantes estimados a partir dos existentes por regressão estocástica.

P: a partir do banco de dados completo formado no passo I são geradas estimativas de médias e matrizes de covariâncias para retroalimentar o passo I.

A partir do passo P a etapa de inserção se completa gerando n bancos de dados completos e únicos. A etapa seguinte (análise) realiza a análise estatística de cada banco de dados gerado na etapa de inserção. E a etapa de combinação gera um único cenário resultante dos diversos bancos de dados gerados. Rubin (1987) determinou que essa combinação fosse a média aritmética das n imputações realizadas.

A variância da combinação é dividida em duas componentes, pois é composta por duas fontes de flutuação: a variância interna de cada cenário, que representa a variância amostral considerando o banco de dados completo e a variância entre cenários, que mede o quão diferem os *n* cenários gerados.

A principal distinção do MI para outros métodos é que esse propõe restaurar a variabilidade dos dados caso estes formassem um conjunto completo. Porém a MI, assim como outros métodos de imputação, não considera a correlação espacial dos dados. Quando se trata de dados geológicos e geometalúrgicos essa relação precisa ser analisada e reproduzida.

Ren (2007) apresentou a metodologia de Atualização Bayesiana (AB) que é baseada no MI, porém considerando a correlação espacial dos dados entre uma variável com amostragem faltante e *n* variáveis secundárias completas, correlacionadas entre si e colocadas.

O método utiliza estimativa bayesiana, que será detalhada no capítulo 2, na primeira etapa do MI. E baseia-se na construção de uma distribuição condicional onde há valores faltantes. A variável incompleta, faltante, é considerada como variável primária e

as outras variáveis completas são consideradas como secundárias, o conjunto é considerado como multivariado Gaussiano. A estimativa Bayesiana se baseia em três distribuições:

- primária, encontrada pelas equações de krigagem simples da variável faltante;
- secundária, distribuição de probabilidade encontrada pela coestimativa com as secundárias;
- posterior, que nada mais é que uma combinação das anteriores.

A completude dos valores faltantes é obtida através de amostragem aleatória por simulação estocástica, após a definição da distribuição posterior. Esse método evoluiu em relação aos demais por considerar a correlação espacial dos dados, que é fundamental quando se trata de dados de um depósito mineral.

As metodologias descritas acima foram desenvolvidas para mecanismo de falta do tipo MAR. Para utilização em mecanismo de falta MNAR, faltante não aleatório, que é o caso do trabalho em questão, é necessário avaliar a característica da distribuição dos dados faltantes para tratá-la de forma adequada.

Rubin (1976) propõe para avaliar dados com mecanismos de falta MNAR, que a probabilidade de falta deve ser baseada em uma variável indicadora de falta R, se o dado for faltante (R=0) ou se for presente (R=1) e criou metodologias sobre esse indicador. O modelo de seleção e o modelo de mistura de padrões, os quais incorporam a propensão de falta de dados de maneiras bastante distintas. O modelo de seleção incorpora uma equação de regressão que prevê a probabilidade de falta. E o modelo de mistura de padrões estratifica o conjunto de dados em padrões de falta e estima os parâmetros individualmente para cada padrão. Porém, os mesmos não são usuais visto que não são passíveis de teste.

Assim, em 1987 Rubin propõe que, sabendo que o mecanismo de falta do conjunto de dados em estudo seja MNAR e que os modelos resultantes do processo de imputação por MLE, MI ou AB são do tipo MAR, deve-se aplicar a esses modelos MAR resultantes uma transformação fixa aos valores imputados com intuito de transformá-los MNAR. O autor sugere adição de 20% ao valor imputado. Já, Cohen (1988) sugere que a transformação fixa seja uma constante de metade do desvio padrão do conjunto completo. Silva (2018) apresenta uma transformação que aplica uma correção igual ao erro máximo obtido.

O trabalho em questão utiliza a transformação fixa dos modelos resultantes MAR em MNAR baseada nos trabalhos listados acima e considera como constante o valor da distância *inter-quartil* da distribuição do erro dos dados após serem complementados.

Com o banco de dados completo, é possível transformá-lo da escala laboratorial (amostra) para a escala operacional (modelo) através da modelagem estatística multivariada (regressão) para previsão geometalúrgica. Essa regressão carrega a não linearidade entre a resposta metalúrgica e o dado químico ou geológico, e essa não linearidade deve ser trazida ao modelamento. Para isso é possível utilizar as seguintes metodologias geoestatísticas: simulação, análise de componentes principais (PCA), função de transferência, teorema de Bayes, krigagem e cokrigagem de indicadores, regressão linear e esperança condicional alternativa ACE.

Segundo Barnett & Deutsch (2013), Nguyen Cong & Rode (1995) e Wang & Murphy (2004), a esperança condicional alternativa (ACE) é um algoritmo de regressão não linear, que maximiza a correlação entre a variável resposta (metalúrgica) e a soma das variáveis preditoras (teores químicos). A geração de um modelo geometalúrgico por regressão ACE tem grande vantagem sobre as outras metodologias listadas por sua habilidade de recuperar a forma operacional das variáveis e de revelar relações complicadas (Wang & Murphy, 2004).

Para gerar um modelo geometalúrgico a partir de um banco de dados onde a variável metalúrgica em questão é faltante do tipo MNAR, é necessário que esta seja completada através de metodologia de inserção. Deseja-se que os bancos de dados completos gerados sejam representativos do fenômeno MNAR. E que finalmente, com os bancos de dados completos sejam desenvolvidos modelos geometalúrgicos respeitando a não aditividade dos dados.

1.3 OBJETIVO

O objetivo deste trabalho é estudar o efeito sobre o modelo de regressão multivariado de um banco de dados, no qual a variável metalúrgica recuperação mássica (RMTOT) possui dados faltantes não aleatórios em relação às demais variáveis de teores químicos (P_2O_5ap, SiO_2, MgO, Fe_2O_3, Al_2O_3, CaO, TiO_2 e RCP).

Pretende-se, ainda, avaliar o possível ganho de acuracidade e precisão do modelo de regressão gerado, utilizando os bancos de dados complementados por imputação em relação ao modelo de regressão gerado ignorando a incompletude do banco de dados.

1.4 METODOLOGIA

Para avaliar o ganho de acuracidade e precisão no modelo geometalúrgico de uma mina de fosfato quando se tem um banco com dado completo em detrimento a um banco com dados faltantes não aleatórios, serão desenvolvidas duas etapas. A primeira é relativa ao tratamento e correção dos dados faltantes por inserção e pode ser sequenciada nas seguintes sub-etapas:

i. O banco de dados inicial com amostras faltantes será separado em dois domínios geológicos: Isalterito de base (ISAB) e Rocha Semi Intemperizada (RSI). Esses domínios são referentes ao grau de alteração do material e são distintos espacialmente.

ii. Os dois bancos de dados com as variáveis completas: P_2O_5, MgO, CaO, Fe_2O_3, Al_2O_3, SiO_2, TiO_2 e RCP e a variável faltante: RMTOT serão avaliados por estatística univariada e multivariada, a fim de determinar a possível eliminação de variáveis redundantes.

iii. Para determinar o valor da transformação fixa que será aplicada ao modelo resultante da inserção, dos bancos de dados isotópico (desconsiderando as amostras faltantes), dos dois domínios, serão removidas 20% aleatoriamente, que serão complementados por meio da atualização Bayesiana e confrontados aos valores originais (omitidos). A taxa de transformação a ser aplicada nos bancos de dados resultantes da próxima etapa, será a relação *inter-quartil* (IQR) do erro relativo dessa inserção.

iv. Passa-se para a etapa de imputação por atualização Bayesiana nos dois bancos de dados com RMTOT faltantes não aleatórios. O método é um procedimento de imputação múltipla que considera a correlação espacial entre as variáveis de acordo com as distribuições primária, probabilidade e posterior. Gera-se os bancos de dados completos em mecanismo MAR.

v. A etapa seguinte será a aplicação da etapa (iii) em (iv), onde os bancos de dados em mecanismo MAR obtidos na etapa (iv) serão transformados em MNAR aplicando a transformação fixa de subtrair o valor da relação *inter-quartil* (IQR) do erro relativo encontrado na etapa iii. Nessa etapa, serão investigados os resultados da escolha da relação IQR como a constante de transformação em detrimento aos demais métodos existentes na literatura para isso.

Com os bancos de dados completados, poderá iniciar-se a segunda etapa que consiste no desenvolvimento do modelo geometalúrgico. Essa pode ser dividida nos seguintes sub-itens:

i. Os domínios geológicos anteriormente separados serão agrupados, visto que na operação de mina os mesmos não são individualizados. Assim, não é possível rastrear essa individualização na alimentação da usina de beneficiamento.

ii. O modelo geometalúrgico será implementado pela metodologia de regressão não linear ACE, tanto para os bancos de dados finais completos como para o banco de dados original do trabalho com RMTOT faltante não aleatório.

iii. Na etapa de reconciliação, os modelos gerados serão comparados aos resultados da usina de beneficiamento a partir de duas bases reconciliáveis. A primeira aplicando-se às pilhas de homogeneização que alimentaram a usina de beneficiamento no ano de 2017, e a segunda aplicando às novas amostras laboratoriais a fim de medir possíveis influências de descontinuidades operacionais da usina de beneficiamento. E por fim, será avaliado os resultados obtidos e conclusões.

1.5 ESTRUTURA DA DISSERTAÇÃO

A dissertação está organizada da seguinte forma: o capítulo 2 aborda uma revisão bibliográfica dos temas (i) flotação e geometalurgia, (ii) metodologias de análise de dados faltantes, (iii) metodologias de imputação, (iv) métodos para geração de modelo geometalúrgico e (v) reconciliação de resultados.

O capítulo 3 apresenta o estudo com o banco de dados real de uma mina a céu aberto de fosfato com a aplicação das etapas listadas no item *1.4 Metodologia*. Além de apresentar os resultados obtidos, avalia-se o ganho de precisão e acuracidade no modelo geometalúrgico. O capítulo 4 traz as conclusões e as recomendações futuras.

CAPÍTULO 2 REVISÕES BIBLIOGRÁFICAS

Este capítulo apresenta os principais fundamentos sobre geometalurgia e flotação, além de uma revisão teórica dos métodos de inserção de dados e dos métodos de geração de modelo geometalúrgico.

2.1 GEOMETALURGIA

As diversas respostas dos minérios frente a um processo são chamadas de variáveis metalúrgicas, as quais podem ser de ordem quantitativas (recuperação metalúrgica e recuperação mássica), qualitativas (teores e propriedades de produtos finais), de impacto ao meio ambiente e de aspecto econômico (consumo específico), dentre outras (Rodriguez *et al*., 1990). A associação dessas variáveis à sua localização é definida com geometalurgia.

O entendimento da geometalurgia de um corpo de minério quantifica o impacto da geologia na moagem, na resposta metalúrgica e no processo de recuperação de metais (Williams & Richardson, 2004). Ou seja, a geometalurgia é a integração dos conhecimentos de geologia, mineralogia, metalurgia e geoestatística, para assim obter um melhor entendimento do depósito, levando ao desenvolvimento de um modelo confiável e robusto, reduzindo riscos e gerando ganhos financeiros.

2.2 FLOTAÇÃO

A flotação foi desenvolvida em 1906 visando o aproveitamento econômico de minérios de baixo teor (Wills, 2007). É um processo, metalúrgico, de concentração de minério através da separação físico-química dos minerais de valor da ganga. Esse processo se baseia na interação de três fases: liquida, sólida e gasosa. O entendimento do processo parte do conceito, que será detalhado a diante, que moléculas polares somente interagem entre si, assim como as moléculas apolares interagem apenas com moléculas apolares.

A fase líquida geralmente é constituída pela água, onde ocorre dissociação das partículas dissolvidas, hidratação e adsorção de íons. A água apresenta alta polaridade, caracterizando-se por ser um forte solvente, assim toda molécula polar é hidrofílica e possui afinidade com a água. A fase sólida é constituída pelo material que será separado (ganga e mineral minério) e é caracterizada pelas propriedades hidrofílicas e hidrofóbicas dos minerais. A fase gasosa é representada pelas bolhas de ar, que são formadas pela injeção do ar na polpa (constituída pela água e pelos minerais a serem separados). Nessa

etapa as partículas hidrofóbicas da polpa se aderem a bolha de ar e são carregadas para a superfície da célula de flotação separando então os minerais polares dos apolares.

A interação entre a bolha de ar e a partícula a ser flotada é um fator determinante na flotação. E necessita que a partícula mineral esteja com a superfície liberada, o que é obtido com a moagem. Segundo Monte & Peres (2004) após a moagem o tamanho ideal da partícula é entre 1mm e 5 micrometros, pois partículas grandes têm dificuldade de aderir à bolha, sendo perdidas durante o deslocamento da bolha até a superfície e partículas muito pequenas não aderem as bolhas sendo perdidas no rejeito.

O grau de hidrofobicidade do mineral na flotação é determinado pela energia de interface entre a bolha de ar e a partícula que é obtida pelo ângulo θ entre elas. Segundo Monte & Peres (2004) quanto maior o ângulo de contato da bolha com a partícula, maior a hidrofobicidade desta e mais fácil é o seu arraste até a superfície. A figura 1 ilustra o ângulo de contato θ.

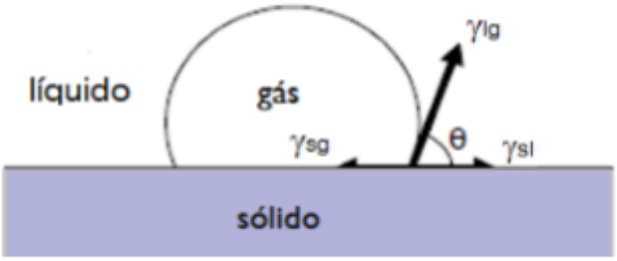


Figura 1 - Ângulo de contato entre bolha de ar e superfície da partícula sólida

Poucos minerais não naturalmente hidrofóbicos, listado por Wills (2007) os seguintes: grafita, enxofre, molibdenita, diamante, carvão, ouro nativo e talco. Para a flotação de minerais naturalmente polares é necessário modificar a superfície do mesmo tornando-a apolar por meio de adição de reagentes ao sistema. De forma que pode-se induzir a hidrofobicidade apenas no mineral de interesse mantendo os outros hidrofílicos, fazendo uma hidrofobicidade seletiva (Monte & Peres, 2004).

2.2.1 Reagentes

Os reagentes são compostos orgânicos e inorgânicos que tem a função de controlar a polaridade das partículas mantendo as características ideais da espuma. Segundo Baltar (2010), os reagentes são classificados de acordo com sua função no processo. Os mais

comuns são os coletores, depressores, espumantes, ativadores, reguladores de pH e dispersantes. Sendo os três primeiros os mais importantes ao sistema.

2.2.1.1 Coletores

O coletor é uma substância que atua na interface sólido-líquido tornando a superfície de um mineral hidrofóbica. Quando ele é adicionado ao sistema, a superfície do mineral adsorve seus os íons ou suas moléculas, tornando-se hidrofóbica, repelente a água. Os coletores são constituídos por uma parte apolar, não iônica e outra polar, iônica. A parte polar da molécula do coletor se liga superfície do mineral e a parte apolar fica em contato com a água, gerando a hidrofobicidade da superfície do mineral possibilitando sua flotação. E para a melhor eficiência dos coletores, estes, após serem inseridos no sistema, devem passar por um período de condicionamento.

Os coletores podem ser divididos em três categorias: catiônicos, aniônicos e não iônicos. A figura 2 ilustra os coletores mais usados na flotação industrial e suas fórmulas químicas estruturais.

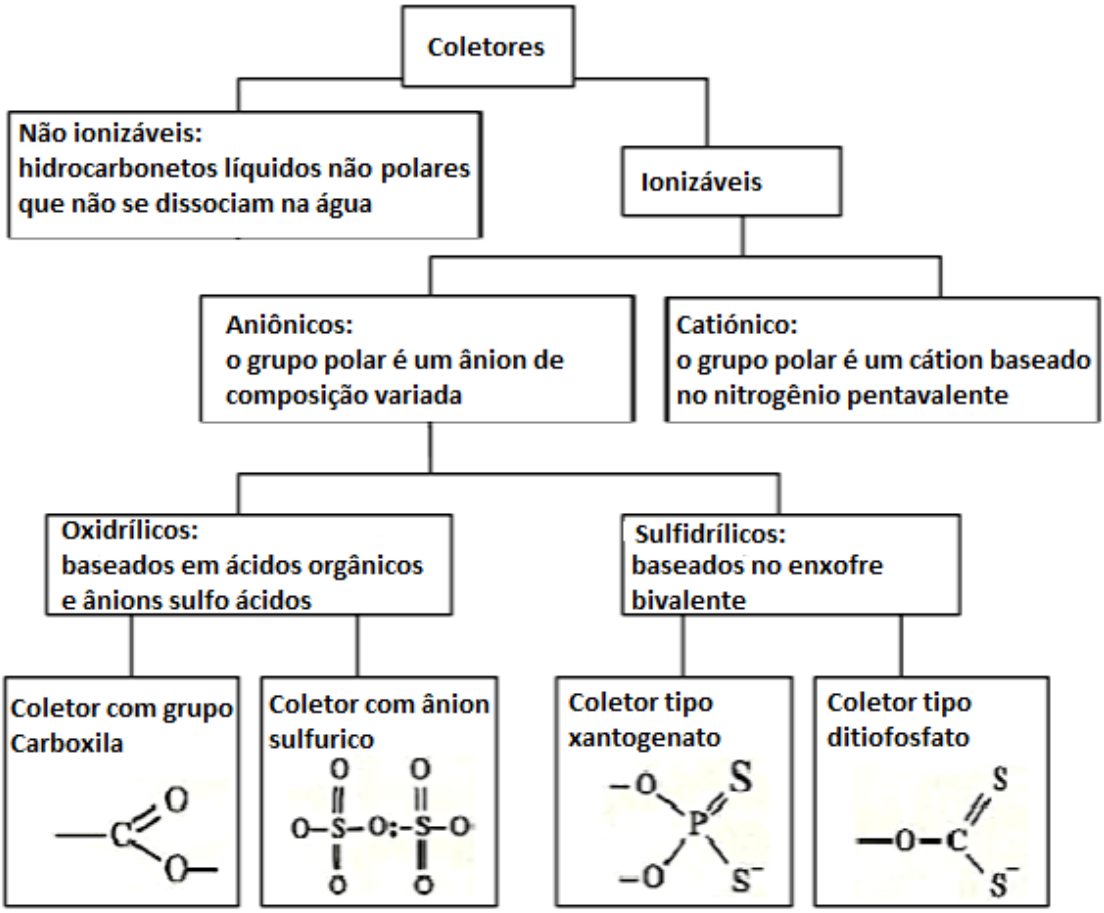


Figura 2 - classificação dos coletores típicos na flotação industrial

Modificada de Nguyen & Schulze (2004).

Os coletores não ionizáveis são substâncias químicas não solúveis em água, utilizados na flotação de minerais naturalmente hidrofóbicos. São exemplos os hidrocarbonetos obtidos do óleo bruto e do carvão (Bulatovic, 2007), como óleo diesel e querosene.

Os coletores aniônicos classificados como sulfidrílicos ou oxidrílico. Os sulfidrílicos são muito empregados na flotação de sulfetos (Rao, 2004), sendo os xantatos os representantes principais. São caracterizados pela ligação de um átomo de enxofre a um átomo de carbono ou fósforo. Já os oxidrílicos têm como característica principal um átomo de oxigênio associado ao carbono ligado à cadeia hidrocarbônica, representam a maior parte dos coletores aniônicos, porém é pequeno o número destes utilizados na flotação industrial (Bulatovic, 2007). Os sabões de ácidos graxos (carboxilas) são oxidrílicos utilizados como coletores de apatita na produção de rocha fosfática e são também empregados como coletores de minerais oxidados e não metálicos (Monte & Peres, 2004).

Os coletores catiônicos são representados pelas aminas e acetonas, se caracterizam por possuir um grupo polar carregado positivamente e um grupo apolar. Segundo Vieira (2016), são muito utilizadas na flotação de sulfetos, silicatos, talco, micas, alguns óxidos de metais raros e para separação de silvita de halita. E as aminas ainda são empregadas na flotação catiônica reversa de minério de ferro e são fortes coletores da apatita podendo flotar seletivamente fosfatos sedimentares de minérios calcários (Wills, 2007).

2.2.1.2 Espumantes

Os espumantes são compostos não iônicos que contém um grupo polar e uma cadeia hirocarbônica capazes de se adsorver na interface líquido-gasosa (Bulatovic, 2007). Tem como função aumentar a estabilidade da espuma, pois reduzem a tensão superficial da água permitindo a sustentação do mineral na bolha até etapa de separação. Os espumantes podem ser sintéticos ou naturais, entre os sintéticos o mais utilizado é o metilisobutilcarbinol (MIBC) e entre os naturais são o óleo de pinho e o ácido cresílico.

2.2.1.3 Modificadores

Os modificadores são substâncias orgânicas e inorgânicas utilizadas para melhorar a seletividade dos minerais na flotação e são classificados de acordo com o papel que

desempenha na flotação. Os modificadores que facilitam a adsorção do coletor em um dado mineral são chamados ativadores (Leja, 1982). Os que agem no sentido de impedir a adsorção do coletor, dificultando a flotação de determinado mineral são chamados depressores (Leja, 1982). Os que modificam o pH do meio são conhecidos como reguladores de pH, e podem ser ácidos ou básicos, dependendo da função modificadora. Os modificadores com função de dispersar as partículas em dimensões coloidais são chamados dispersantes. E são usados para o controle de pH e muitas vezes agem como depressores de ganga (Peres & Araújo, 2006).

2.2.2 Circuito de beneficiamento do fosfato

O minério de fosfato pode ter origem ígnea ou sedimentar e sua origem afeta o processo de beneficiamento devido aos diferentes minerais de ganga presente. Dentre os métodos de beneficiamento de fosfato a flotação é o mais utilizado mundialmente, como cerca de 80% dos depósitos brasileiros, como o do estudo em questão, são de origem ígnea, a flotação será detalhada para esse tipo de depósito.

As plantas de beneficiamento de fosfatos são compostas pelas etapas de cominuição, classificação, deslamagem, separação magnética e flotação direta da apatita em pH básico por uma rota que contém as etapas de *rougher, scavenger e cleaner* e que podem apresentar diferentes circuitos de acordo com a granulometria do minério (Leal Filho & Chaves, 2004). São utilizadas para o sistema de flotação células mecânicas e/ou colunas de flotação. De acordo com Fernandes (2013) a última vem ganhando espaço nos últimos anos devido à obtenção de melhores resultados.

Os reagentes utilizados na flotação de fosfato são coletores depressores e modificadores de pH. Segundo Guimarães *et al.* (2005) os coletores mais utilizados na flotação de fosfatos ígneos são os ácidos graxos, saponificados com NaOH, a fim de produzir sais solúveis para atuar como coletores da apatita. Os ácidos graxos mais comuns são os ácidos oleico e linoleico derivados de espécies vegetais como o óleo de casca de arroz e o óleo de soja. E o depressor utilizado em todos os sistemas de flotação nacional é o amido de milho, por apresentar o melhor resultado dentre todas as substâncias testadas até hoje.

2.2.3 Avaliação dos resultados da flotação

A avaliação do desempenho da flotação pode ser feita pela relação da recuperação mássica, recuperação metalúrgica e o teor do elemento de interesse no concentrado. Para calcular as recuperações considera-se a massa da alimentação (massa do ROM) que é expressa pela equação 1:

$$Massa\ do\ ROM = massa\ do\ concentrado + massa\ do\ rejeito \quad (1)$$

E a recuperação mássica (RMTOT), como:

$$RMTOT = \frac{massa\ do\ concentrado}{massa\ do\ ROM}\ x\ 100\% \quad (2)$$

E a recuperação metalúrgica (RECMET) é a relação da recuperação mássica com os teores de alimentação e do concentrado, como pode ser visto na equação 3.

$$RECMET = \frac{teor\ \%\ no\ concentrado}{teor\ \%\ no\ ROM}\ x\ RMTOT \quad (3)$$

É importante considerar que tanto a recuperação mássica como metalúrgica podem ser afetas por vários fatores na flotação, como: qualidade da água, presença de lama, geometria da coluna de flotação, vazão da entrada de ar, tipo de reagente, tempo de condicionamento, granulométrica das partículas, dentre outros (Persechini *et al.*, 2001). O número e distribuição das análises do parâmetro de recuperação frente ao das análises químicas ditam o grau de conhecimento geometalúrgico do depósito.

2.3 METODOLOGIAS DE ANÁLISE DE DADOS FALTANTES E IMPUTAÇÃO DE DADOS

Em mineração para que o empreendimento mineral possa ser planejado de forma adequada é necessário o desenvolvimento de um modelo que represente os fenômenos geológicos do depósito, para que esse modelo não carregue viés é necessário conhecer todos os valores em todos os locais amostrados. Porém nem sempre todos são conhecidos, assim a geoestatística vem desenvolvendo métodos para prever os valores faltantes, os quais serão abordados neste item.

2.3.1 Estimativa por Máxima Verossimilhança (Maximum Likelihood Estimation – MLE)

Segundo Enders (2010) a estimativa por máxima verossimilhança baseia-se na distribuição da população dos dados, sendo ela uma distribuição normal conhecida por

sua média e variância. E pode ser descrita pela distribuição de probabilidade do conjunto de dados através da função de densidade de probabilidade, dada por:

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

Sendo z_i um valor amostral e μ a média da população, σ^2 a variância da população, assim G_i é a probabilidade relativa de se obter um determinado valor amostral a partir da distribuição normal dos dados.

A equação 4 dá a probabilidade de se obter um valor amostral, porém o objetivo é obter um conjunto de valores amostras. Segundo (Enders, 2010) isto se obtém por meio do produto das probabilidades individuais, ou seja, a probabilidade conjunta de eventos independentes ocorrerem. Os resultados desse produto são de pequena magnitude possibilitando aplicar-se o logaritmo natural a esses valores sem perder significado.

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

Assim, aplicando-se a equação acima é possível estimar os valores faltantes da população por um processo iterativo que aplica diversos valores de média e variância para encontrar o melhor ajuste da população de dados, ou seja, que maximize o logaritmo da probabilidade conjunta dos dados.

A descrição apresentada é atribuída ao caso univariado, quando o objetivo é aplicar a um conjunto multivariado essa aplicação é direta. Da mesma forma, se busca a distribuição normal multivariada que melhor ajuste o conjunto amostral, pela função de densidade de probabilidade multivariada:

$$G_i = \frac{1}{(2\pi)^{n/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-0.5(\boldsymbol{\Phi}_i-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Phi}_i-\boldsymbol{\mu})} \qquad (6)$$

Sendo μ como o vetor de médias, $\sum$ a matriz de covariâncias das variáveis, Φ_io vetor randômico de observações e n o número de amostras pertencentes a cada variável.

Como no caso univariado, o logaritmo da função densidade de probabilidade multivariada do conjunto amostral é o somatório das funções densidade de probabilidade, ou seja,

$$\log G = \sum_{i=1}^{N} \log G_i \qquad (7)$$

A aplicação do método de estimativa por máxima verossimilhança em análise de dados faltantes quando o mecanismo de falta é MAR é recomendado por Enders (2010),

visto que ela produz estimativas sem enviesamento. Para isso, Dempster *et al*. (1977) desenvolveu o algoritmo EM. Cada interação do algoritmo envolve duas etapas:

- A etapa E (esperança) preenche os dados faltantes a partir das amostras completas e da construção de um conjunto de equações de regressão estocástica baseadas no vetor de médias, na matriz de covariância.
- A etapa M (maximização) utiliza o conjunto completo obtido na etapa E e encontra um novo vetor de médias e uma nova matriz de covariâncias.

A partir da etapa M o algoritmo realiza uma nova interação de E com os novos valores do vetor de médias e da matriz de covariância e assim por diante. Essas interações se repetem até que seja atingida a convergência, obtendo o vetor de médias e a matriz de covariância que maximizam a função de densidade de probabilidade multivariada (Enders, 2010).

O algoritmo EM é um procedimento que realiza a estimativa dos parâmetros (vetor de médias e a matriz de covariância) em conjuntos amostrais incompletos, e também pode ser utilizado como uma ferramenta para inserção de dados. Porém essa estimativa por máxima verossimilhança só deve ser aplicada quando o mecanismo de falta é MAR, pois caso o mecanismo de falta seja MNAR essa metodologia irá produzir estimativas enviesadas.

2.3.2 Estimativa bayesiana

A estimativa bayesiana é um método estatístico que combina crença e probabilidade, consiste em estimar os parâmetros de uma distribuição (posterior), que descreve a probabilidade relativa para diferentes valores de parâmetros. A distinção entre estimativa bayesiana e a abordagem frequentista, utilizada em testes em disciplinas de psicologia, economia, etc, está na definição de um parâmetro. Na abordagem frequentista um parâmetro é definido como um valor desconhecido, porém fixo que assume o valor que maximiza a função densidade de probabilidade dos dados observados. Já a abordagem bayesiana define os parâmetros como variáveis aleatórias que possuem uma distribuição. Um dos objetivos de uma análise bayesiana é descrever a forma dessa distribuição.

A análise Bayesiana consiste em três etapas principais: (1) especificar uma distribuição prévia para o parâmetro de interesse, (2) usar uma função de verossimilhança para sumarizar a evidência dos dados sobre diferentes valores de parâmetros e (3) combinar informações da distribuição anterior (1) e a função densidade de probabilidade

(2) para gerar uma distribuição posterior que descreva a probabilidade relativa de diferentes valores de parâmetros. Descrever a forma da distribuição posterior é o objetivo principal da análise bayesiana (Enders, 2010).

A primeira etapa (1) consiste em especificar uma distribuição prévia baseada no que se acredita sobre a probabilidade relativa de valores diferentes dos parâmetros antes mesmo de coletar quaisquer dados. Assim, nessa etapa é necessário definir os parâmetros da distribuição a priori, como: a média a priori, variância dessa distribuição e um valor suposto de pontos amostrais coletados.

A segunda etapa (2) consiste em sumarizar a evidência dos dados sobre diferentes valores de parâmetros por meio da função de densidade de probabilidade. O que significa que nessa etapa, são coletados dados e testado um conjunto de parâmetros para descrever a distribuição e retornar as probabilidades relativas de cada parâmetro.

A etapa final da análise bayesiana é definir a distribuição posterior (3), que é composta pelas informações da função prévia e a função de probabilidade, pesando cada ponto na função de verossimilhança pela magnitude da função prévia.

Por trás da análise bayesiana, há o teorema de Bayes. Esse descreve a relação entre duas probabilidades condicionais. Para dois eventos aleatórios, A e B:

$$p(B|A) = \frac{p(B)p(A|B)}{p(A)} \qquad (8)$$

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

Substituindo na equação 8, A pelo dado amostral (X) e B por um parâmetro (θ), tem-se:

$$p(\theta|X) = \frac{p(\theta)p(X|\theta)}{p(X)} \qquad (9)$$

Assim os termos da equação 9 estão alinhados com os conceitos da estimativa bayesiana. Onde, θ é o parâmetro de interesse, X é a amostra, $p(\theta)$ é a distribuição prévia do parâmetro, $p(X|\theta)$ é função probabilidade, $p(X)$ é a distribuição marginal dos dados, e $p(\theta|X)$ é a distribuição posterior que se busca na estimativa bayesiana, ou seja, a probabilidade condicional do parâmetro θ, dado o que é sabido dos dados.

A distribuição posterior nada mais é do que uma função de verossimilhança ponderada, onde a ideia básica é ajustar cada ponto na função de verossimilhança pela magnitude da distribuição prévia. Isso é realizado no numerador do teorema de Bayes, multiplicando a função de verossimilhança pela distribuição a priori. O denominador do

teorema nada mais é do que uma constante de escala que faz a área sob a distribuição posterior ser igual a um. Assim ao dividir por uma constante não alteraria a forma básica da distribuição posterior, portanto, ignorar o denominador produz a seguinte expressão simplificada:

$$Posterior \propto Priori \times função\ de\ verossimilhança$$

Desta forma o teorema de Bayes permite a atualização das informações sobre os parâmetros da distribuição à medida que novas amostras são coletadas. O que torna a análise bayesiana a base para a imputação múltipla, quando aplicada ao vetor média e a uma matriz de covariância de dados multivariados.

2.3.3 Inserção múltipla

A imputação múltipla uma alternativa a estimativa por máxima verossimilhança, pois a estimativa de máxima verossimilhança estima os parâmetros diretamente a partir dos dados disponíveis e a inserção múltipla realiza a estimativa dos parâmetros da distribuição dos dados, preenchendo os valores ausentes antes da análise. Assim, gera n cenários completos distintos com n desvios padrões. Possibilitando prever a incerteza dos valores imputados. A inserção múltipla (MI) (Rubin, 1987) é realizada em três etapas: inserção, análise e combinação.

2.3.3.1 Inserção

A etapa da inserção gera várias cópias do banco de dados e preenche os valores faltantes de cada cópia com diferentes estimativas. Este processo é subdivido em dois passos: I e P e usa um algoritmo iterativo que repete em ciclos o passo de inserção (I) e o passo posterior (P). O passo I utiliza equações de regressão estocástica para inserir os valores faltantes, e a etapa P usa os dados preenchidos para gerar novas estimativas do vetor média e da matriz de covariância.

A geração de múltiplos conjuntos de valores imputados em cada passo I requer estimativas diferentes do vetor média e da matriz de covariância, que são o objetivo do passo P. Em cada etapa P, o algoritmo iterativo usa os dados preenchidos do passo I anterior para definir as distribuições posteriores do vetor média e da matriz de covariância. Em seguida, ele usa a simulação de Monte Carlo para "desenhar" novas estimativas do vetor média e da matriz de covariância de seus respectivos posteriores. O próximo passo I usa esses valores atualizados de parâmetros para construir um novo conjunto de equações de regressão que são ligeiramente diferentes daquelas no passo I

anterior. Repetindo o procedimento das duas etapas várias vezes geram várias cópias do bando de dados, cada uma das quais com estimativas únicas dos valores faltantes (Enders,2010).

Todos os aspectos da inserção múltipla estão enraizados na Metodologia Bayesiana, mas o passo P está particularmente ligado à estimativa bayesiana, pois busca descrever a distribuição posterior do vetor médio e da matriz de covariância.

- O passo I os valores faltantes são completados pela seleção aleatória em uma distribuição preditiva, que depende dos valores observados x_{obs}, e das estimativas do vetor média e da matriz de covariância. A equação de inserção é:

$$x_i^* \sim p(x_{mis}|x_{obs}; \mu_{i-1}^*) \qquad (10)$$

sendo x_i^* o valor estimado no instante i, x_{mis}o valor faltante, x_{obs}, os valores observados e μ_{i-1}^*o vetor de média e matriz de covariâncias calculados no instante anterior a i.

- O passo P desenha aleatoriamente um novo vetor média e uma nova matriz de covariância a partir de suas respectivas distribuições posteriores. Essas novas estimativas são geradas por simulação computacional de Monte Carlo. O passo P usa os dados preenchidos no passo I precedente para calcular um vetor de médias amostrais e uma matriz de somas quadráticas e produtos cruzados. A partir das quais é possível obter a distribuição posterior da matriz de covariância e o vetor de médias atualizado. A descrição completa das equações posteriores do vetor média e da matriz covariância é encontrada em Enders (2010).

Depois de calculados e atualizados o vetor de médias e a matriz de covariância o algoritmo retorna ao passo I para calcular um novo conjunto de regressões e gerar um novo banco de dados completado. Os passos I e P são repetidos n vezes permitindo gerar n bancos de dados completos. De forma sucinta o passo P pode ser descrito pela seguinte equação:

$$\Theta_i^* \sim p(\Theta | x_{obs,} x_i^*) \qquad (11)$$

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

2.3.3.2 Análise e combinação

A etapa de análise tem o objetivo de avaliar a estatística dos parâmetros e estimativa de cada um dos n cenários gerados a etapa anterior. A partir das n estatísticas

geradas a etapa de combinação tem o objetivo de consolidar os n cenários em um único final. A forma de combinação é definida por Rubin (1987) como a média aritmética das estimativas pontuais de cara realização de inserção. Como pode ser visto na equação:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i^*}{n} \qquad (12)$$

onde x_i^* é a i-ésima estimativa pontual e $\bar{x}$ é a estimativa pontual combinada, que de acordo com a abordagem bayesiana é a média da distribuição posterior dos dados observados.

Rubin (1987) define que a variância da MI possui duas fontes de flutuação: a variância interna à inserção e a variância entre inserções. Que são respectivamente, o erro amostral caso do banco completado e o erro amostral resultado dos dados faltantes. Sendo a variância interna à inserção σ_I^2:

$$\sigma_I^2 = \sum_{i=1}^{n} \frac{\sigma_i^2}{n} \qquad (13)$$

onde σ_i^2 é a variância do i-ésimo cenário e n é o número de cenários. Já a variância entre inserções σ_E^2 é dada por:

$$\sigma_E^2 = \sum_{i=1}^{n} \frac{(x_i^* - \bar{x})^2}{n-1} \qquad (14)$$

onde σ_E^2 representa a variabilidade dos parâmetros entre os diversos cenários. Assim a variância total da MI (σ_T^2) pode ser dada pela combinação de σ_I^2 e σ_E^2:

$$\sigma_T^2 = \sigma_I^2 + \sigma_E^2 + \frac{\sigma_E^2}{n} \qquad (15)$$

Onde $\frac{\sigma_E^2}{n}$ é um fator de correção sabendo-se que o número de realizações é finito.

2.3.4 Estimativa bayesiana aplicada a dados correlacionados

As técnicas listadas até o momento não consideram a correlação espacial entre os dados. A consideração da correlação espacial quando se trata de um depósito mineral é fundamental para que este seja reproduzido espacialmente de forma coerente. A atualização bayesiana desenvolvida por Doyen *et al*. (1996) e Ren (2007) propõe considerar a correlação espacial entre as amostras e sua teoria é desenvolvida abaixo.

Considere uma variável aleatória X estacionária dentro de uma área A como sendo a variável primária de interesse. E um conjunto de m variáveis aleatórias $Z_j, j = 1 \ldots m$ como sendo as variáveis secundárias. E assumindo que X e Z_j são multigaussianas.

Ren (2007) utiliza os resultados de krigagem simples da variável de interesse X como parâmetro da distribuição primária da estimativa bayesiana. A qual, como descrita

anteriormente, é composta por três distribuições: primária, função de probabilidade e posterior. Assim:

$$\bar{X}_p(u) = \sum_{i=1}^{n} \lambda_i . X_p\,(u_i) \qquad (16)$$

Onde $\overline{\mathrm{X}}_{\mathrm{p}}$ é a média da distribuição primária, uo local de estimativa, u_i o local onde a variável de interessa está amostrada e λ_i os pesos de krigagem simples que são encontrados por:

$$\sum_{i=1}^{n} C(u_i - u_k) = C(u - u_k); k = 1, \dots, k \qquad (17)$$

$\mathrm{C}(\mathrm{u_i} - \mathrm{u_k})$ é a covariância entre a variável primária no local $\mathrm{u_i}$ e no local $\mathrm{u_k}$, e $\mathrm{C}(\mathrm{u} - \mathrm{u_k})$ a covariância entre a primária no local a ser estimado e no local $\mathrm{u_k}$. E a variância da distribuição é obtida por:

$$\sigma_p^2(u) = 1 - \sum_{i=1}^{n} \lambda_i C(u, u_i) \qquad (18)$$

A partir das equações 17 e 19 dos parâmetros a distribuição primária gaussiana está definida.

Se os dados secundários Z_j estão disponíveis em todos os locais da área, são secundários colocados fornecendo outra distribuição condicional. Esta distribuição é a função de probabilidade da estimativa bayesiana e é definida por:

$$\bar{X}_L(u) = \sum_{i=1}^{m} \lambda_i . Z_i(u) \qquad (19)$$

em que $\mathrm{Z_i(u)}$ é o valor da amostra secundária no local de estimativa u e $\overline{\mathrm{X}}_{\mathrm{L}}$ a média da função densidade de probabilidade. Os pesos são calculados pela equação 20.

$$\sum_{j=1}^{m} \lambda_i . \rho_{j,k} = \rho_{i,0} \qquad (20)$$

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

$$\sigma_L^2(u) = 1 - \sum_{j=1}^{m} \lambda_i . \rho_{i,0} \qquad (21)$$

Com as distribuições primaria e de função probabilidade definidas pode-se calcular a posterior pelo produto das duas primeiras. A descrição desse produto, obtendo as expressões de média e variância da distribuição posterior, é apresentada em Ren (2007) e Silva (2018), finalizando assim, a atualização bayesiana para dados correlacionados.

Segundo Barnett & Deutsch (2012) pela amostragem da distribuição posterior é possível realizar a imputação de dados por simulação estocástica, mantendo constantes as amostras existentes no banco de dados e obtendo a incerteza dos valores imputados pela distribuição posterior. Para um melhor entendimento do funcionamento da atualização bayesiana Ren (2007) representou-a em forma de figura:

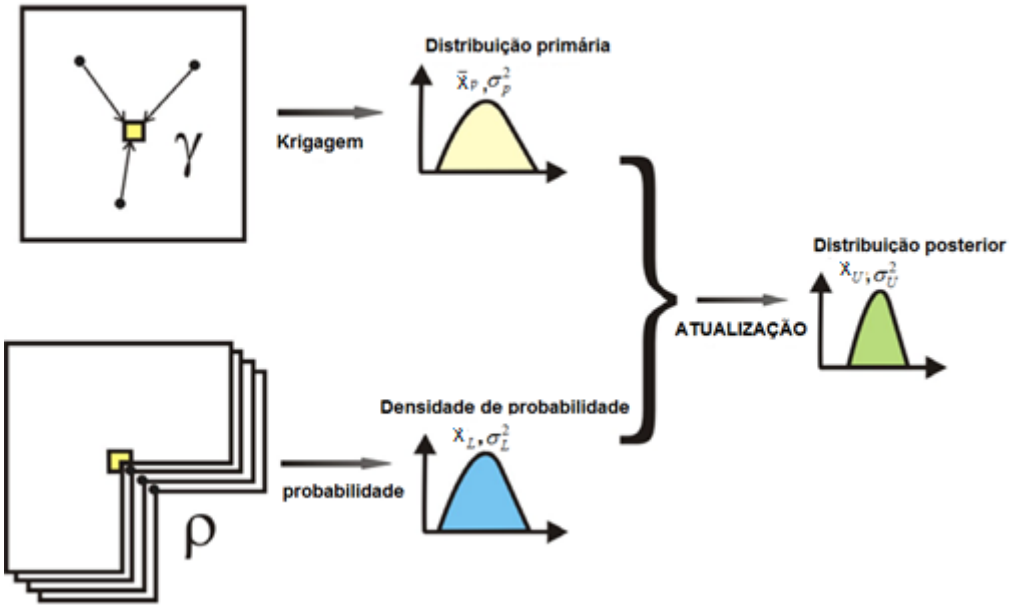


Figura 3 - Diagrama da atualização Bayesiana. O quadrado amarelo representa o local a ser complementado e os pontos pretos as amostras disponíveis. Figura adaptada (Ren, 2007)

Um ponto importante a ser considerado na atualização bayesiana é que para a utilização desta é necessário que os dados estejam em ambiente multigaussiano. Como dados geológicos reais não tem suas distribuições condicionais e marginais Gaussianas tem que aplicar alguma técnica para transformá-las em gaussinas. A transformação *nscore* (Verly, 1984; Deutsch & Journel, 1998) pode ser aplica aos dados originais transformando-os em gaussianos, porém deve-se aplicar a retrotransformação às distribuições resultantes da atualização Bayesiana para trazê-los ao ambiente original dos dados.

2.3.5 Metodologias MNAR

Os métodos de análise tratados até agora assumem a falta de dados aleatórios (MAR). Embora as abordagens baseadas em MAR representem o atual estado da arte (Schafer & Graham, 2002), uma quantidade considerável de pesquisa metodológica é dedicada modelos para dados ausentes não aleatoriamente (MNAR). Sabendo que os valores não amostrados carregam informações da distribuição dos parâmetros da população, as metodologias para tratar dados MNAR devem considerar tais informações para não acarretar em estimativas enviesadas e inacuradas desses parâmetros. Dentre as metodologias para tratar dados MNAR estão: o modelo de seleção e o modelo de mistura de padrões. Ambos os modelos descrevem a distribuição conjunta dos dados e a

probabilidade de falta, mas de maneira distinta. O modelo de seleção considera uma equação de regressão adicional que prevê as probabilidades de falta de dados. Em contraste, o modelo de mistura de padrões forma subgrupos de casos que compartilham o mesmo padrão de dados faltante e estima o modelo de análise dentro de cada padrão.

Alguns pontos sobre tais modelos são importantes de serem considerados: o fato de o modelo de seleção depender fortemente de premissas não testáveis sobre a distribuição e o fato de que o modelo de mistura de padrões exige que se especifique valores assumidos para um ou mais parâmetros inestimáveis. Sobre tais aspectos não há como verificar se esses requisitos são atendidos e se violações podem produzir estimativas distorcidas, tal fato levou alguns autores a alertar contra o uso das mesmas (Allison, 2002; Demirtas & Schafer, 2003; Schafer & Graham, 2002).

2.3.5.1 Modelo de Seleção

O modelo de seleção (Heckmann, 1979) parte da ideia de considerar que cada variável do conjunto de dados possui um par de amostras: Z e R. Onde Z é um valor que pode ou não ser observado e R é um código binário que representa o fato da variável ter sido amostrada ou não. Assim, R=1 se Z foi amostrado e R=0 se Z é amostra faltante (Rubin, 1976). Quando o mecanismo de falta é MNAR, se sabe que o modelo que define o fato de R assumir o valor um ou zero é desconhecido, assim os dados amostrados e faltantes são dependentes entre si e sua distribuição é dada por $p(Z, R)$. O modelo de seleção propõe que essa distribuição conjunta seja fatorada em duas distribuições componentes:

$$p(Z, R) = p(R|Z)p(Z) \qquad (22)$$

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

Heckmann (1979) trata essa fatoração em duas partes, combinando um modelo de regressão, chamado de fundamental, pois é o modelo que seria estimado caso o conjunto de dados fosse completo. E um modelo de regressão que prevê a probabilidade de resposta de uma variável. Ou seja, define a probabilidade de ocorrerem faltantes como uma variável latente, com distribuição normal. O modelo de regressão fundamental, que corresponde a distribuição marginal de Z, seria:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (23)$$

onde β_0e β_1 são os coeficientes de regressão, y a variável faltante, x a variável correlacionada a y e ε o termo residual de regressão. O modelo de regressão que prevê a probabilidade de falta seria:

$$F = \alpha_0 + \alpha_1 w + \delta \qquad (24)$$

onde α_0 e α_1 são os coeficientes de regressão, w a variável completa correlacionada a x e y, e δ o termo residual. Esta equação corresponde à distribuição condicional de falta. Como F é uma completamente faltante, é usada então a variável binária R que descreve se o valor é presente ou não no conjunto de dados. Para construir um modelo de regressão em que a variável de saída é binária pode-se utilizar o modelo *probit* (Bliss, 1934).

O modelo *probit* usa a transformação de distribuição normal padrão cumulativa para gerar previsões de probabilidades. Sendo a distribuição normal uma função em forma de S, cuja altura corresponde à proporção da curva normal padrão que cai abaixo de um determinado valor z. Assim esse modelo expressa a probabilidade prevista de uma resposta completa como:

$$p(R = 1|w) = \phi[\alpha_0 + \alpha_1 w] \qquad (25)$$

onde ϕ representa a distribuição cumulativa normal padrão e α_0 e $\alpha_1$os coeficientes de regressão. A estimativa de $\phi[\alpha_0 + \alpha_1 w]$ resulta no valor de probabilidade de falta que se busca estimar.

Desta forma se tem a distribuição marginal dos dados (equação 23) e a distribuição condicional de falta (equação 25) que definem o modelo de seleção, e assim a distribuição conjunta dos dados completos e faltantes. Através de sorteios de valores desta distribuição conjunta por simulação de monte carlo, é obtido a imputação de dados completando os mesmos.

Enders (2010) destacou que a utilização do modelo de seleção é limitada, pois erros amostrais na estimativa podem ser gerados, quando houver correlação entre as variáveis preditoras do modelo *probit* e do modelo fundamental acarretando em correlação entre as probabilidades estimadas e às variáveis do modelo fundamental. Além disso, caso a distribuição dos dados não seja normal o modelo produzirá estimativas distorcidas. E de maneira prática, não tem como avaliar a qualidade do modelo de seleção, pois para que tenha desempenho satisfatório é necessário que todas as hipóteses assumidas por este sejam cumpridas e essas apresentam alta instabilidade.

2.3.5.2 Modelo de mistura de padrões

O modelo de mistura de padrões integra a distribuição que descreve os dados faltantes na análise estatística. Especificamente, a abordagem de mistura de padrões forma subgrupos de casos que compartilham o mesmo padrão de falta e estima a distribuição dentro de cada subgrupo, cada subgrupo será chamado de padrão de seleção. Voltando à fatoração na Equação 22, observe que o modelo de mistura de padrões fatora de maneira distinta ao modelo de seleção, cada modelo de padrão específico corresponde à distribuição condicional, p(Z|R) e as proporções de cada padrão de falta correspondem a distribuição marginal, p (R). De maneira que a distribuição de Z está condicionada ao mecanismo de falta R:

$$p(Z, R) = p(Z|R)p(Z) \qquad (26)$$

Os padrões de falta podem ser classificados como monotônicos, onde as variáveis estão dispostas em uma ordem tal que, se a variável é faltante para Z_j é faltante também para Z_{j+1}. Ou não monotônico, onde a falta não é ordenada entre variáveis.

O modelo de mistura de padrões também apresenta limitações, pois é preciso que atribua valores a parâmetros que em essência são inestimáveis, devido a incompletude do conjunto de dados. Caso os valores atribuídos estejam incorretos este pode gerar valores ainda mais enviesados que os obtidos por meio de modelo MAR (Enders, 2010).

2.3.5.3 Transformações fixas

Devido às limitações listadas acima e, até recentemente, à falta de opções de software a aplicação dos modelos de seleção e de mistura de padrões ficou pouco usual. Assim Rubin (1987) apresentou a transformação fixa que lida de forma prática com modelo de falta MNAR. Para o autor, no modelo de falta MNAR a diferença entre os dados observados e os dados faltantes é sistemática, o que impossibilita a estimativa direta do viés gerado por essa diferença amostral, assim o modelo gerado será sensível às hipóteses consideradas em relação às similaridades entre observados e não observados. Ou seja, o modelo gerado representará a relação assumida entre os dados completos e faltantes.

Bancos de dados com mecanismo de falta MNAR, onde a probabilidade de uma amostra ser faltante depende do seu valor, carregam viés, o qual é sempre incorporado aos resultados de estimativa por máxima verossimilhança ou por imputação múltipla e

suas derivações. Já que estas metodologias assumem mecanismo MAR. Assim para possibilitar a utilização dessas metodologias sem incorporação de viés é necessário acrescentar aos dados uma variável que descreva a probabilidade de falta no mecanismo MNAR em questão.

A metodologia de Rubin (1987) para tratativa dessa sensibilidade consiste em gerar múltiplas imputações sob um mecanismo MAR, depois adicionar uma constante aos valores imputados tornando-os saída MNAR. Estimados os cenários completos por máxima verossimilhança, imputação múltipla, ou por atualização bayesiana, assumindo mecanismo MAR, adiciona-se uma constante aos resultados sub ou superestimados. De forma que essa constante possa baixar a superestimativa ou aumentar a subestimativa. Rubin (1987) recomenda que a cada resultado imputado baseado em mecanismo MAR seja adicionado um valor constante que aumenta ou diminui os resultados das imputações em 20%, transformando-o em MNAR. Porém essa sugestão é arbitrária.

Uma abordagem alternativa é a de Cohen (1988) que sugere utilizar como constante o desvio padrão dos cenários complementados que tem como vantagem a relação com a variabilidade do conjunto de dados e a métrica familiar.

Silva (2018) propôs uma abordagem baseada em Rubin (1987) na qual a transformação fixa se dá pela adição de uma constante igual ao erro relativo máximo obtido pelo mecanismo MAR.

$$Err_j = \frac{z(u_i) - z_j^*(u_i)}{z(u_i)} \qquad (27)$$

$$z_{j'}^*(u_i) = z_j^*(u_i) + z_j^*(u_i) * Err_j \qquad (28)$$

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

Considerando as metodologias apresentadas o trabalho em questão baseou-se na abordagem de Silva (2018) do erro relativo. Porém utilizou como constante a relação *inter quantil* da distribuição do erro relativo e não o erro máximo, com o intuito de desconsiderar valores anômalos dos erros encontrados.

O cálculo do erro relativo é realizado segundo a metodologia de Silva (2018), a qual do banco de dados original são removidas 20% das amostras da variável de interesse, aleatoriamente para não gerar viés aos dados em análise. O conjunto de dados resultante, chamado de conjunto de calibragem, é submetido à imputação por atualização bayesiana sob mecanismo MAR. Assim pela comparação entre os valores originais que foram

removidos e os valores complementados obtém-se diretamente o erro relativo do processo de inserção. O número de complementações (cenários) realizadas ao conjunto de calibragem deve ser o mesmo a ser realizado no banco de dados original com dados faltantes não aleatórios.

Após o cálculo dos erros relativos é possível encontrar a relação interquartil das suas distribuições em cada cenário, como exemplificado na figura abaixo.

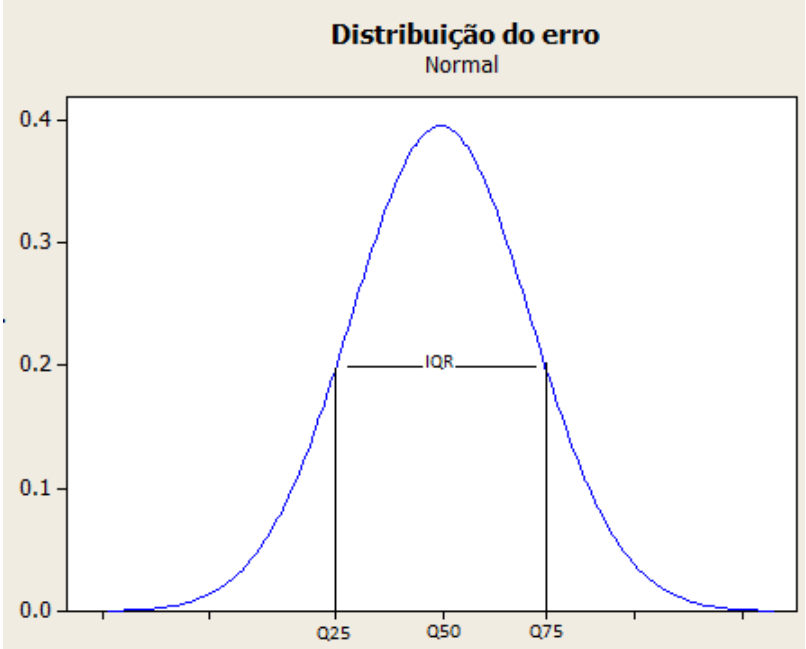


Figura 4 - Distribuição do erro relativo, mostrando a distância interquartil (IQR)

A relação interquartil (IQR) é calculada como a distância entre o quartil 25% e o quartil 75% da distribuição e mostra o espalhamento da distribuição, desconsiderando a influência de valores extremos. Assim a IQR do erro de cada cenário obtido será subtraída dos valores complementados proporcionalmente no cenário correspondente. Por exemplo, considerando que a IQR do cenário 1 seja 0.27, aos valores complementados no cenário 1 será subtraído 27% de seu valor. Ou seja, se $z_i^*(u_i) = 12.3\%$, o valor MNAR depois de aplicada a transformação fixa será $z_{i'}^*(u_i) = 12.3 - (12.3 * 0.27)\% = 8.979\%$, considerando que o viés é devido à ausência de valores baixos. Caso o viés for gerado por ausência de valores altos o teor corrigido será 15.621%.

2.4 METODOLOGIAS DE ESTIMATIVA DE MODELO GEOMETALÚRGICO

A partir de um banco de dados complementado, onde tanto as variáveis químicas quanto a variável metalúrgica em questão têm resultados em todos os pontos de amostragem passa-se para a etapa de desenvolvimento do modelo geometalúrgico que represente os fenômenos geológicos do depósito, para que este possa ser utilizado no planejamento de lavra para previsão da recuperação do bem mineral no processo de beneficiamento.

2.4.1 Krigagem de indicadores

A krigagem foi desenvolvida por Matheron no início da década de 60, a qual consiste em diversos métodos de estimativa que visa minimizar a variância do erro. Os métodos de krigagem podem mais usuais são: a krigagem simples, ordinária e de indicadores (Deustch & Journel, 1998; Sinclair & Blackwell, 2004).

Segundo Isaaks e Srivastava (1989) o estimador de krigagem ponderado de acordo com o espaçamento das amostras e o agrupamento das mesmas, quanto mais próximas elas estiverem do ponto a ser estimado maior o seu peso no estimador, e quanto mais agrupadas menor o peso individual.

A krigagem de indicador (KI), introduzida por Journel (1983), é um método muito utilizado na estimava de dados não lineares, para lidar com valores extremos e com diferentes padrões de continuidade. Ela define a probabilidade de um determinado valor ocorrer no local em análise e pode ser aplicada para variáveis contínuas, como variável metalúrgica, e variáveis categóricas, como domínio geológico.

A ideia básica da KI é discretizar uma distribuição contínua em K teores de corte, obtendo-se K funções indicadoras. De acordo com Isaaks e Srivastava (1989) o indicador da variável contínua regionalizada $I(u; z_k)$ tem dois possíveis valores, 0 para valores acima de um determinado teor de corte, ou 1 para valores abaixo de teor de corte, como mostra a equação abaixo.

$$i(u_\alpha; z_k) = \begin{cases} 1 \; para \; z(u_\alpha) \leq z_k \\ 0 \; para \; z(u_\alpha) > z_k \end{cases} \quad k = 1, \dots, k \qquad (29)$$

onde $i(u_\alpha; z_k)$ é o indicador da amostra z no local u_α ser maior ou menos que o teor de corte z_k.

A KI requer um variograma baseado no código binário que cada dado recebe com relação a um teor de corte z_k, ou seja, um variograma para cada categoria, ou classe de teor (Journel, 1983).

A krigagem de indicadores de uma variável aleatória fornece a estimativa da função de distribuição acumulada condicional (ccdf) para um teor de corte z_k. Os dados do atributo contínuo z são discretizados em k classes, e para cada classe é calculada a proporção dos z-dados que não excede um determinado teor de corte. A ccdf, construída a partir da krigagem de indicadores das k classes, representa um modelo de probabilidade da incerteza sobre os valores não amostrados $z(u)$ (Deustch & Journel, 1998).

Segundo Goovaerts (1997) o excesso de classes demanda um grande esforço computacional, e escassez de classes causa perda de informações da distribuição. Assim, recomenda-se um número mínimo de 5 classes e máximo de 15, a fim de se obter uma discretização razoável da distribuição e um bom desempenho computacional.

O modelo resultante da krigagem de indicador é representado por uma função de probabilidade binária bloco a bloco de o teor ser acima ou abaixo do teor de corte. Sendo assim os blocos de minério a serem avaliados para determinar seu resultado metalúrgico quando lavrados não apresentam um valor resultante único, mas sim a probabilidade de estarem acima ou abaixo do teor de interesse. O que torna a metodologia pouco usual para o objetivo do trabalho em questão.

2.4.2 Regressão linear

Regressão linear é um método que visa obter a relação entre uma variável dependente ou resposta Y, e uma ou mais variáveis independentes ou preditoras $x_1 \ldots, x_n$, a fim de predizer o valor da variável resposta. A relação entre X e Y é expressa através do diagrama de dispersão, no qual é possível verificar qual o comportamento da variável resposta com as variáveis de entrada, seja ele: linear, quadrático, cúbico, exponencial ou logarítmico. Caso haja correlação entre as variáveis independentes e a variável dependente, determina-se a equação de regressão (Sarma, 2009).

Segundo Sinclair e Blackwell (2004) a correlação mede a similaridade entre os pares de variáveis. E pode ser determinado pelo coeficiente de correlação ρ, que pode ser definido de duas formas: o coeficiente de correlação de Pearson e coeficiente de correlação rank.

O coeficiente de correlação de Pearson pode ser encontrado pela covariância dividida pelos desvios padrões das variáveis:

$$\rho = \frac{1/n \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \quad (30)$$

Onde $1/n \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$é a covariância entre x e y , x_i são os possíveis valores de x e $\bar{x}$ é a média de x. y_i são os possíveis valores de y e $\bar{y}$ é a média de y. E σ_x e $\sigma_y$são, respectivamente, os desvios padrões de x e y. (Waller & Gotway, 2004).

Segundo Issaks e Srivastava (1989) quanto mais próximo de |1| for ρ maior é a correlação entre as variáveis de análise, e quanto mais próximo de 0, menor é a correlação. Nesse caso o diagrama de dispersão será uma nuvem de pontos difusa.

O coeficiente de correlação rank, também conhecido como coeficiente de Spearman, mede a correlação entre duas variáveis x e y quando estras não possuem distribuição normal (Landim, 2003). E pode ser calculado como:

$$\rho_{rank} = \frac{\frac{1}{n}\sum_{i=1}^{n}(R_{x_i} - \overline{R_x})(R_{y_i} - \overline{R_y})}{\sigma_{R_x}\sigma_{R_y}} \quad (31)$$

onde R_{x_i} é a posição de x_i dentre os outros valores de x, R_{y_i} é a posição de y_i dentre os outros valores de y , os quais são calculados classificando-os em ordem crescente. $\overline{R_x}$ é a média das posições $R_{x_1}, \dots, R_{x_n}$; σ_{R_x} é o desvio padrão de R_x; $\overline{R_y}$ é a média das posições $R_{y_1}, \dots, R_{y_n}$; σ_{R_y} é o desvio padrão de R_y.

O coeficiente de correlação de *rank* sofre menos influência de valores extremos que o coeficiente de correlação de Pearson. Um coeficiente de correlação *rank* alto e um coeficiente de Pearson baixo pode indicar a presença de alguns pares erráticos. Semelhantemente, um coeficiente de correlação de Pearson com valor muito alto enquanto o coeficiente de correlação *rank* é baixo, indica a presença de pares de valores extremos (Rossi & Deutsch, 2014).

2.4.2.1 Regressão linear simples

Na regressão linear simples utiliza-se apenas uma variável independente para explicar a variável resposta. Como pode ser descrito pela equação abaixo:

$$y = b_0 + b_1 x + \varepsilon \quad (32)$$

onde x é a variável independente, b_0 é o ponto de intercessão da reta com o eixo Y, b_1 é a inclinação da reta e ε é o erro associado. Essa equação é usada para previsões futuras de y ou para estimar a resposta média de um valor específico de x (Fávero, Belfiore, Silva & Betty, 2009).

Os coeficientes b_0 e b_1 são estimados pelo método de mínimos quadrados de Johnson (2002), que tem como objetivo minimizar a soma do quadrado dos resíduos. As equações abaixo apresentam os cálculos desses coeficientes:

$$b_0 = \bar{y} + b_1\bar{x} \quad (33) \qquad b_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \quad (33)$$

2.4.2.2 Regressão linear múltipla

A regressão linear múltipla envolve duas ou mais variáveis independentes para explicar a variação da variável dependente. Em geral, a adição de variáveis independentes fornece um melhor ajuste da reta e aumenta a correlação entre os dados teóricos e os reais (Sarma, 2009; Wackernagel, 1998). A equação de regressão múltipla é dada por:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n + \varepsilon \quad (34)$$

onde y é a variável resposta, $x_1, \ldots, x_n$são as variáveis preditoras, $b_0, \ldots b_n$, são os coeficientes de regressão e ε é o erro, que possui distribuição normal. A função de regressão múltipla pode ser chamada de superfície de resposta e sua representação gráfica é demostrada na figura 5.

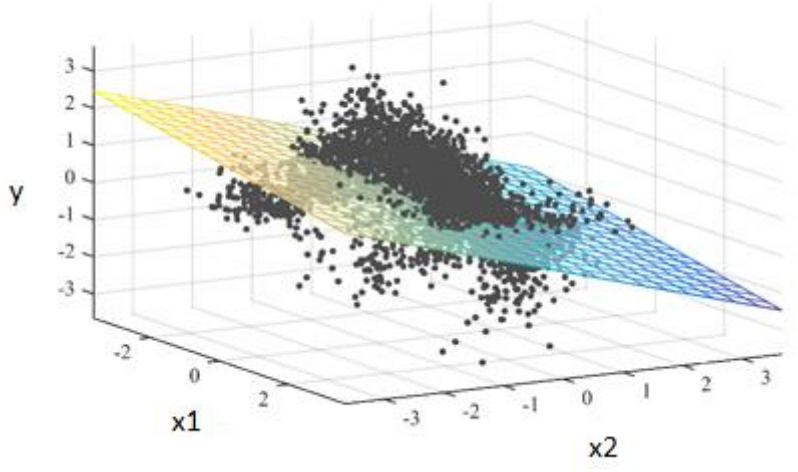


Figura 5 - Representação gráfica de uma função de regressão múltipla

Os coeficientes no modelo de regressão múltipla, assim como no modelo de regressão simples são calculados pelo método de mínimos quadrados. E para tais modelos algumas suposições devem ser assumidas, como:

- Os resíduos devem ter distribuição normal, assim a variável resposta tem uma distribuição normal condicionada às preditoras.
- Os resíduos têm variância constante e ele é assumido como sendo homocedástico condicional às variáveis preditoras. Quando isso não ocorre, é improvável

que os mínimos quadrados proporcionem o melhor ajuste, já que as observações receberão pesos diferentes com base em sua variância.

- As variáveis preditoras devem ser livres de erro, o modelo de regressão assume erro na resposta, mas não pode haver erro associados às variáveis preditoras. Pois isso pode causar que o poder preditivo das variáveis independentes seja subestimado o que é chamado de viés de atenuação.

2.4.2.3 Regressão polinomial

A regressão polinomial é usada quando o modelo apresenta comportamento de polinômio, ou seja, a resposta é curvilínea. Nesse caso, ao invés de ajustar a uma reta, o ajuste é feito por meio de uma função polinomial de grau igual ou superior a dois (Brownlee, 1967).

Esse é um caso especial da regressão linear, onde a função é curvilínea, pois considera $x_{i1} = x_i$ e $x_{ik} = x_i^k$. Assim o modelo $y = xb + \varepsilon$ também é usado para ajustar a regressão polinomial.

A regressão quadrática é um tipo de regressão polinomial representada pela equação:

$$y = b_0 + b_1 x_i + b_2 x_i^2 + \varepsilon \qquad (35)$$

A superfície resposta da regressão quadrática é demonstrada na figura 6.

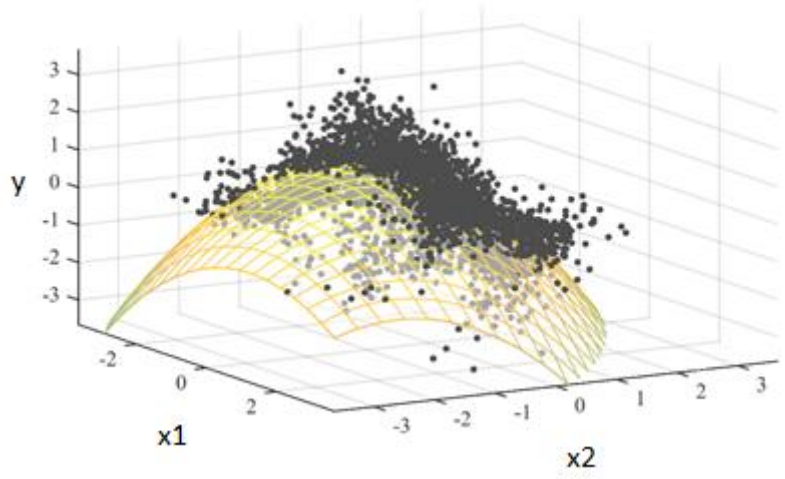


Figura 6 - Representação gráfica de uma função de regressão polinomial

2.4.3 Esperança condicional alternada (*Alternating Conditional Expectation* – ACE)

Regressões paramétricas, como linear e polinomial, fazem suposições fortes e restritivas sobre as relações entre as variáveis preditoras e de resposta. O modelo de regressão linear é uma estrutura muito simples onde toda variância não explicada pela função linear é associada ao erro ε (Barnett & Deutsch, 2013). No entanto, muitas das relações entre variáveis respostas e preditoras que a geoestatística tenta prever não estão relacionadas de maneira linear. Mesmo adotando funções não lineares como a quadrática ou cúbica não é suficiente para descrever relações complexas como as criadas por fenômenos geológicos. Pois essas técnicas tradicionais de regressão são limitadas, já que requerem suposições a priori sobre a relação entre as variáveis preditoras e de resposta (Wang & Murphy, 2004).

Para simplificar e explicar a covariância entre variável resposta e preditoras na análise de regressão é sugerido que as variáveis sejam transformadas de forma adequada. Alguns autores como Box e Cox (1964), Kruskal (1965), Box e Tidwell (1962), Mosteller e Tukey (1977), Cook e Weisberg (1982), Carroll e Ruppert (1988) e Royston (2000) citados por Wang & Murphy (2004) sugeriram métodos de transformações paramétricas para modelar o efeito da covariância. Já outros também citados por Wang & Murphy (2004), como: Royston e Altman (1994), Durrleman e Simon (1989), Hastie e Tibshirani (1990), Green and Silverman (1994) e Bowman e Assalini (1997) desenvolveram técnicas de ajuste de curvas por transformações não paramétricas, motivados pelo objetivo de encontrar ajustes ótimos para a regressão, diminuindo assim a necessidade de suposições sobre a superfície de regressão. Uma abordagem completa sobre esses métodos não paramétricos pode ser encontrada em Härdle (1992).

As regressões não paramétricas, segundo Wang e Murphy (2004), são baseadas no refinamento sucessivo buscando alcançar a superfície de regressão ótima, mantendo-se orientada pelos dados. O método de regressão não paramétrico ACE (Alternating Conditional Expectation) foi desenvolvido por Breiman e Friedman (1985) para estimar transformações ótimas tanto para a variável resposta quanto para as variáveis preditoras na análise de regressão e correlação. A vantagem da abordagem ACE reside na sua capacidade de recuperar as formas funcionais das variáveis e revelar e simplificar as relações complexas entre elas. O modelo geral da regressão ACE é descrito pela equação abaixo:

$$\theta(Y) = \alpha + \sum_{i=1}^{n} \varphi_i(x_i) + \varepsilon \qquad (36)$$

onde θ é a função de transformação da variável resposta Y, e φ_i as funções de transformações das variáveis preditoras $x_i, i = 1, \dots, n$. Assim, o modelo ACE mitiga o problema de estimar uma única função linear de n-dimensional $x = x_1, \dots, x_n$, estimando n funções unidimensionais separadas φ_i e a função θ. Essas funções transformações são estimadas de forma iterativa, buscando minimizar a variância inexplicada da relação linear entre a variável resposta transformada e a soma das variáveis preditoras transformadas.

O algoritmo inicia o processamento gerando transformações para as variáveis preditoras e resposta com média zero e estandardizando $\theta(Y)$ de forma a ter variância igual a 1, ou seja, $E[\theta^2(Y)] = 1$ é possível obter a variância do erro ε^2 da regressão ACE com a equação 37:

$$\varepsilon^2 = E\{[\theta(Y) - \sum_{i=1}^{n} \varphi_i(x_i)]\}^2 \qquad (37)$$

Para a minimização da variância do erro ε^2 o algoritmo de ACE utiliza as equações 38 e 39:

$$\varphi_i(x_i) = E[\theta(Y) - \sum_{j \neq i}^{n} \varphi_j(x_j) | x_i] \qquad (38)$$

onde a otimização da i-nésima transformação de variável preditora $\varphi_i(x_i)$ é obtida pela diferença esperada entre $\theta(Y)$ e $\sum_{j \neq i}^{n} \varphi_j(x_j)$ dentro da condicional de x_i. E a otimização transformação da variável resposta $\theta(Y)$ considerando a transformação das variáveis preditoras constante, é obtida dividindo o valor esperado de $E[\sum_{i=1}^{n} \varphi_i(x_i)]$ pela sua magnitude dentro da condicional de Y (equação 40).

$$\theta(Y) = \frac{E[\sum_{i=1}^{n} \varphi_i(x_i)|Y]}{||E[\sum_{i=1}^{n} \varphi_i(x_i)|Y]||} \qquad (39)$$

A metodologia de Esperança condicional alternativa (ACE) pode ser descrita de forma resumida como sendo a iterativa minimização das funções de esperança condicional, como descrito nas equações 38 e 39.

As transformações das variáveis preditoras $\varphi_i(x_i), i = 1, \dots, n$ e resposta $\theta(Y)$ pós minimização iterativa são estimadas por suas transformações ótimas, assim no ambiente transformado a relação entre variáveis preditoras e resposta se dá por:

$$\theta^*(Y) = \sum_{i=1}^{n} \varphi_i^*(x_i) + e^* \qquad (40)$$

onde e^* é o erro não mitigado pelo uso das transformações ACE e este é assumido como tendo distribuição normal, com média 0 e é obtido pela relação com o máximo coeficiente de correlação múltipla ρ^* da ACE:

$$e^{*2} = 1 - \rho^{*2} \qquad (41)$$

A regressão ACE por suas características de minimização do erro e maximização da correlação entre as variáveis foi a metodologia adotada para geração do modelo geometalúrgico, no qual a variável resposta é a recuperação mássica e as variáveis preditoras os teores químicos. A figura 7 apresenta graficamente a regressão ACE.

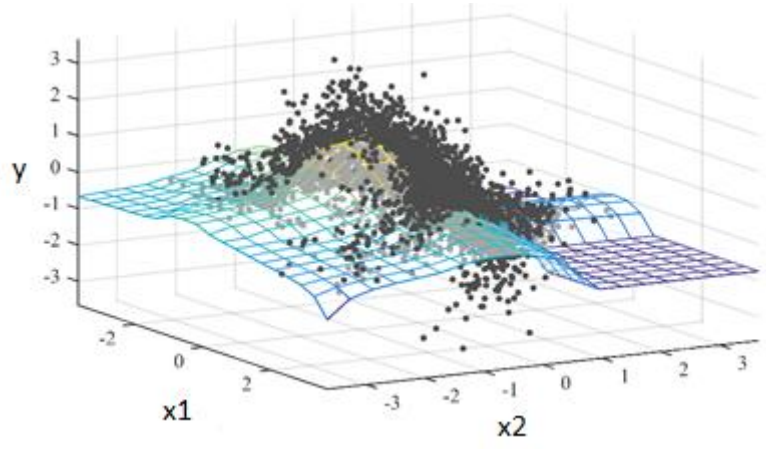


Figura 7 - Representação gráfica da função de regressão ACE

Considerando as metodologias listadas nesse capítulo, foi adotada nesse trabalho, a atualização bayesiana desenvolvida por Doyen *et al.* (1996) e Ren (2007) para complementação dos dados faltantes assumindo mecanismo MAR e a transformação fixa baseada na metodologia de Silva (2018) considerando a transformação pelo valor da distância *inter-quartil* da distribuição do erro e a geração do modelo geometalúrgico por regressão ACE.

2.5 PILHA DE HOMOGENEIZAÇÃO

Para avaliar os resultados obtidos pelas metodologias utilizadas nesse trabalho e descritas anteriormente, a regressão ACE será aplicada sobre pilhas de homogeneização.

As pilhas de homogeneização são desenvolvidas com objetivo de minimizar a variabilidade residual do minério proveniente de lavra, pois quando é bem planejada e operada garante a homogeneização do minério.

De acordo com Schofield (1980) os principais atenuadores da variabilidade do minério, em qualquer sistema de homogeneização em pilhas, são os equipamentos utilizados, a forma de deposição e a retomada do material. Para definir o modelo a ser adotado é necessário conhecer a disponibilidade de espaço do pátio, as distâncias entre o pátio e a usina e entre o pátio e as frentes de lavra, além das características granulométricas do minério. O modelo de pilha adotado na mina em estudo é o longitudinal em Chevron.

As pilhas em Chevron são formadas pela deposição de sucessivas camadas de minério sobrepostas que formam uma espécie de prisma contendo minério de diversas frentes de lavra. A retomada é feita em fatias verticais, de forma que cada fatia é formada por blocos de lavra de áreas diferentes da mina. A figura 08 apresenta o esquema simplificado do modelo de pilha em Chevron.

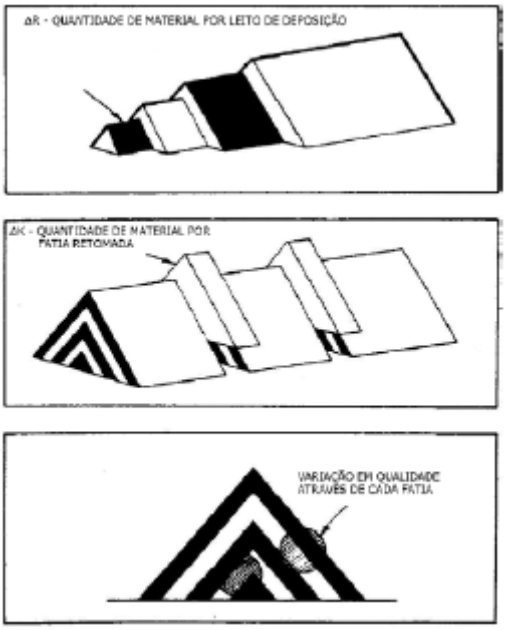


Figura 8 - esquema simplificado do modelo Chevron para pilhas (Schofield, 1980)

CAPÍTULO 3 ESTUDO DE CASO

Este capítulo apresenta o desenvolvimento do modelo geometalúrgico, de uma mina de fosfato, por regressão ACE em um conjunto de dados que será complementado por atualização bayesiana e transformado utilizando a distância *inter-quartil* do erro, como descrito no capítulo 2.

De forma simplificada o estudo de caso desenvolve-se segundo o fluxograma da figura 09:

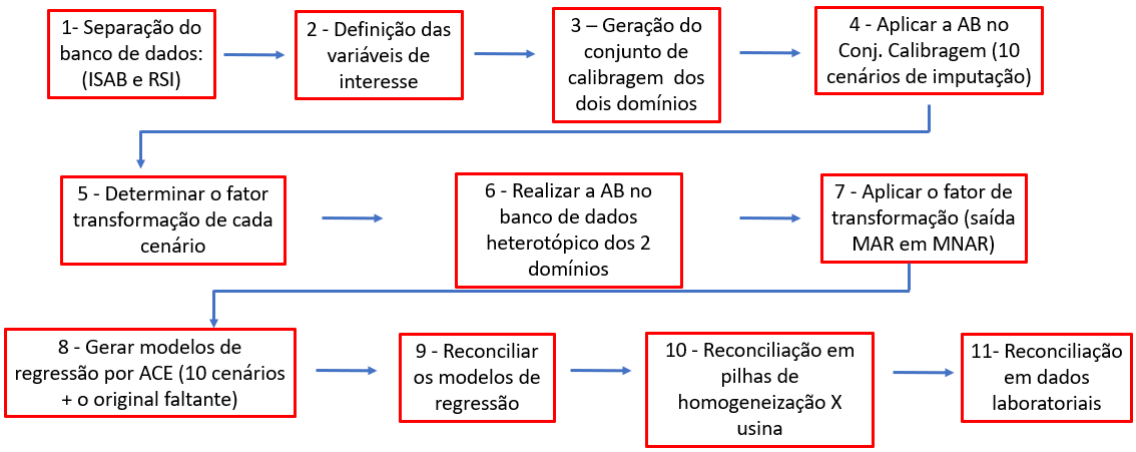


Figura 9 - Fluxograma das etapas desenvolvidas no estudo de caso

3.1 BANCO DE DADOS

O estudo de caso foi realizado em uma mina de fosfato, desenvolvida em um complexo carbonatítico constituído por rochas silicáticas (predominantemente ultrabásicas), carbonatíticas e foscoríticas (Brod *et al*. 2000). A concentração de apatita e anatásio, nesse depósito, é relacionada ao manto de intemperismo que se desenvolveu sobre essas rochas alcalinas. A alteração supergênica destes minerais se deu por solubilização e lixiviação de componentes mais instáveis contidos nas rochas originais. Nos horizontes mais superficiais, a apatita foi parcialmente transformada em fosfato secundário, mas em porções mais profundas permaneceu no manto de intemperismo como um mineral resistato. Os horizontes mineralizados em apatita são o isalterito de base (ISAB) e rocha semi intemperizada (RSI), que são ilustrados na figura 8:

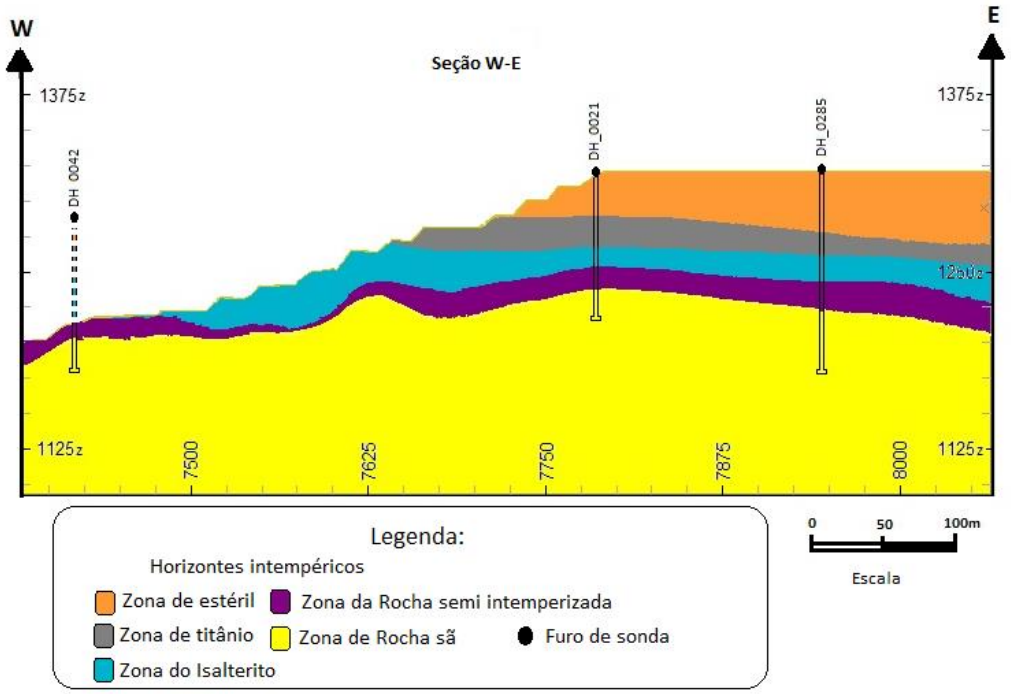


Figura 10 - Perfil intempérico do depósito. Figura adaptada de Leal, et al. (2016)

O banco de dados do estudo é constituído por amostras de ISAB e RSI, coletadas em uma malha de aproximadamente 25 x 25 x 10m, e enviadas ao laboratório químico para obter os teores de P_2O_5ap, SiO_2, MgO, Fe_2O_3, Al_2O_3, CaO, TiO_2 e RCP, sendo RCP a razão entre CaO e P_2O_5 a qual determina a porcentagem do P_2O_5 associado à apatita. Essa determinação é feita da seguinte forma:

- Se RCP≥1,35, P_2O_5ap=P_2O_5, se não P_2O_5ap=CaO/1,35.

Os teores de P_2O_5ap e RCP são definidores do *cutoff* do intervalo de minério. São consideradas como minério apenas as amostras com:

- Teor de P_2O_5ap ≥ 5,
- E 0,9≤ RCP ≤3.

Assim, apenas as amostras dentro dos limites listados acima, são consideradas como mineralizadas e são enviadas à planta piloto para determinação do potencial de recuperação de massa na usina de beneficiamento, ou seja, do teor de recuperação mássica (RMTOT). A recuperação mássica, que é a variável de interesse a ser estimada no modelo geometalúrgico, e medida com base na massa da alimentação (massa do ROM) e massa do concentrado, como descrito nas equações 1 e 2 no capítulo 2.

O banco de dados em questão é constituído por 3221 amostras de ISAB e 1749 de RSI como distribuição representada pela figura 11. Das 3221 amostras de ISAB, somente

3004 foram enviadas à planta piloto e possuem resultado de RMTOT, e das 1749 amostras de RSI apenas 1113, esses números estão apresentados na tabela 1. Desta forma, o banco de dados é heterotópico. A figura 12 ilustra a distribuição espacial do banco evidenciando as amostras com dados faltantes, ou seja, os que não possuem resultados de RMTOT. Nota-se, que as amostras incompletas estão distribuídas de forma difusa; porém, esses 217 dados incompletos do ISAB e 636 do RSI são referentes às amostras com teores fora da especificação do minério, para as quais é esperado em planta piloto resultados baixos de recuperação mássica. Os resultados faltantes ocorrem devido ao próprio valor amostral, ou seja, apenas nas amostras de baixo teor do mineral minério, portanto diz-se que os dados metalúrgicos são faltantes não aleatórios. Apesar de serem amostras consideradas como estéril, essas regiões ocorrem de forma difusa na mina e muitas vezes acabam sendo lavradas, em baixa proporção, junto à massa de minério. Os histogramas das figuras 13 e 14 ilustram a distribuição dos dados evidenciando em vermelho a região da curva onde faltam amostras.

Tabela 1 - Síntese do quantitativo de amostras com e sem resultado de RMTOT.

N. de amostras	Domínio		
	ISAB	RSI	Total
Com resultado de RMTOT	3004	1113	**4117**
Sem resultado de RMTOT	217	636	**853**

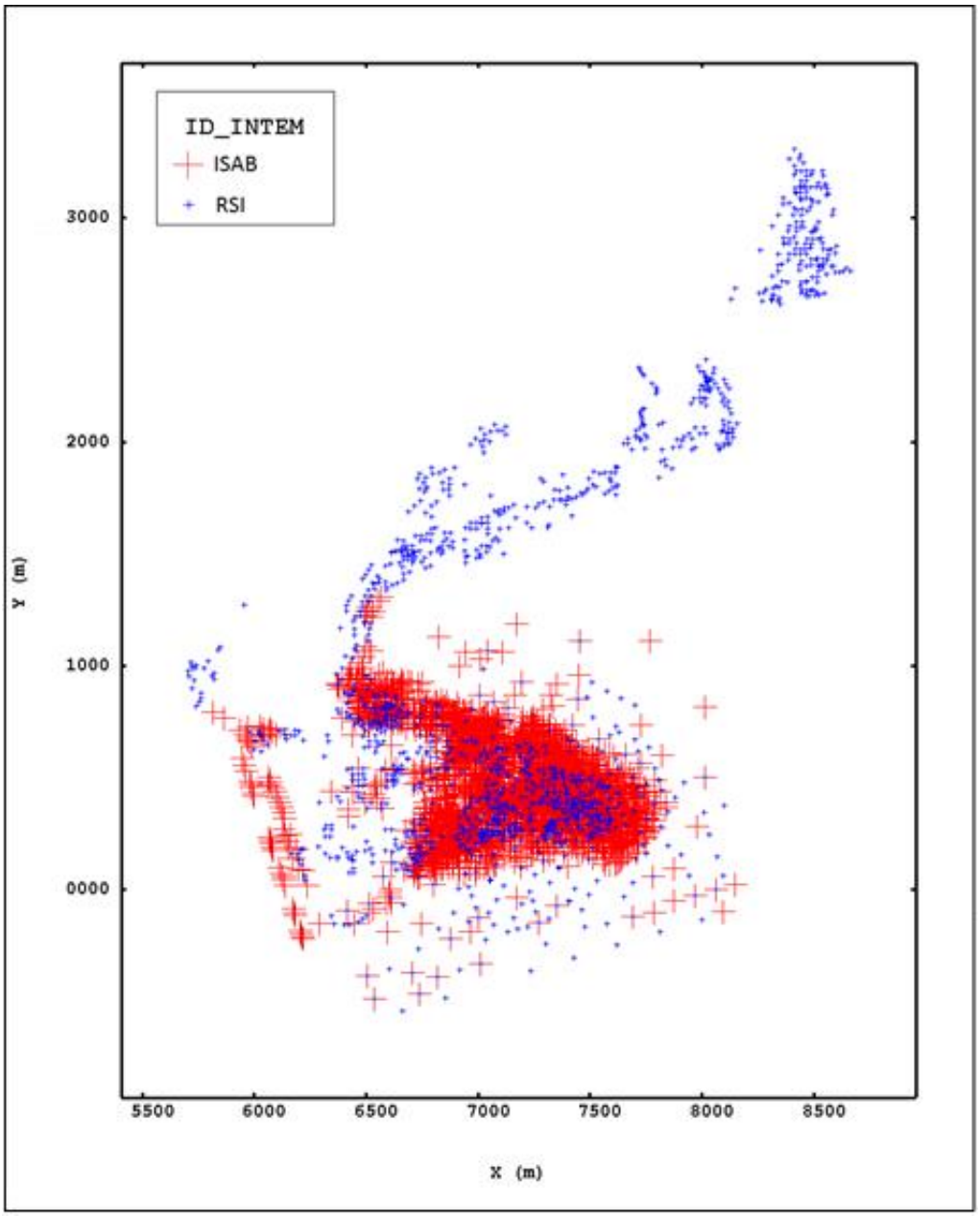


Figura 11 - Distribuição espacial das amostras por domínio

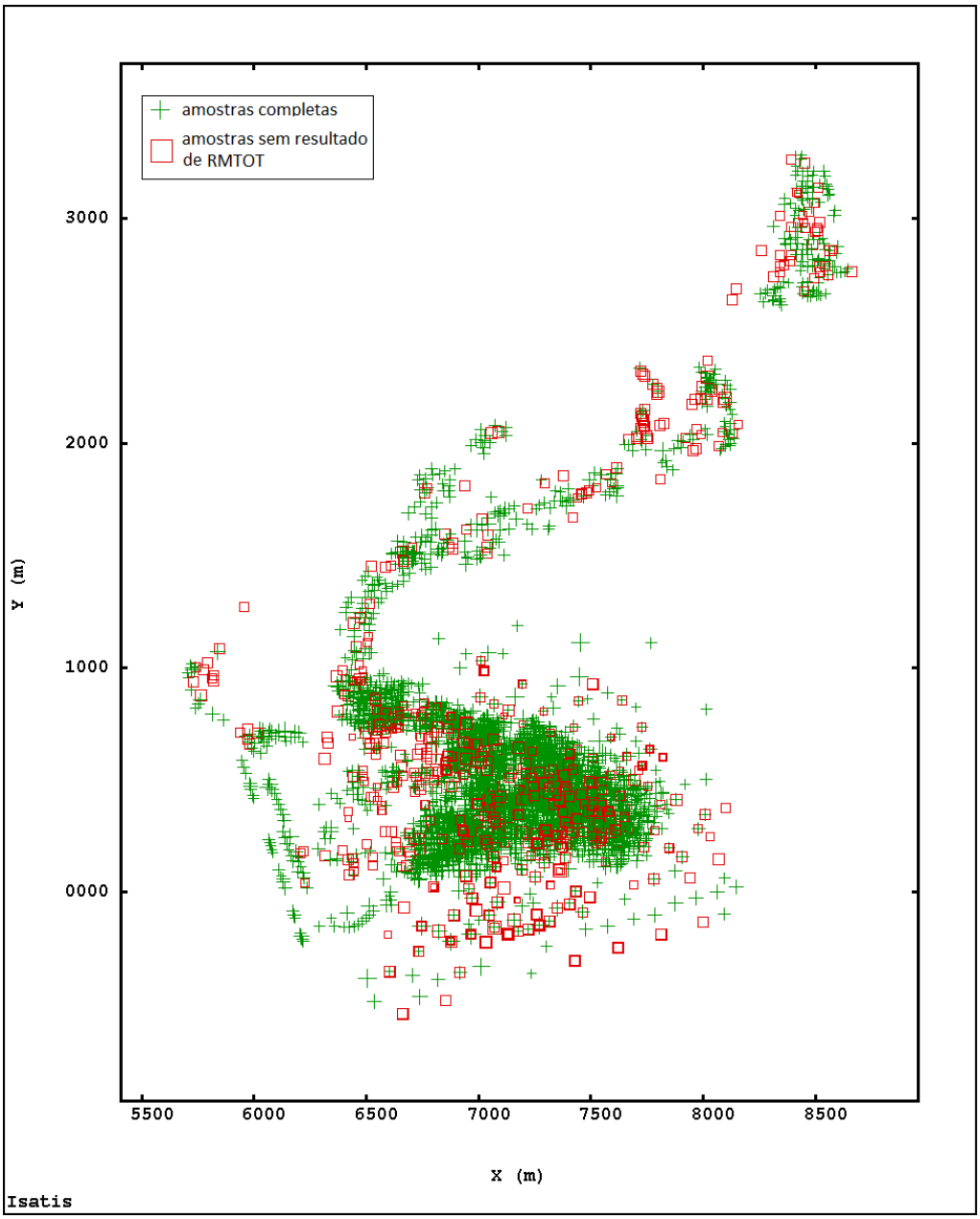


Figura 12 - Distribuição espacial das amostras evidenciando as sem resultado de RMTOT

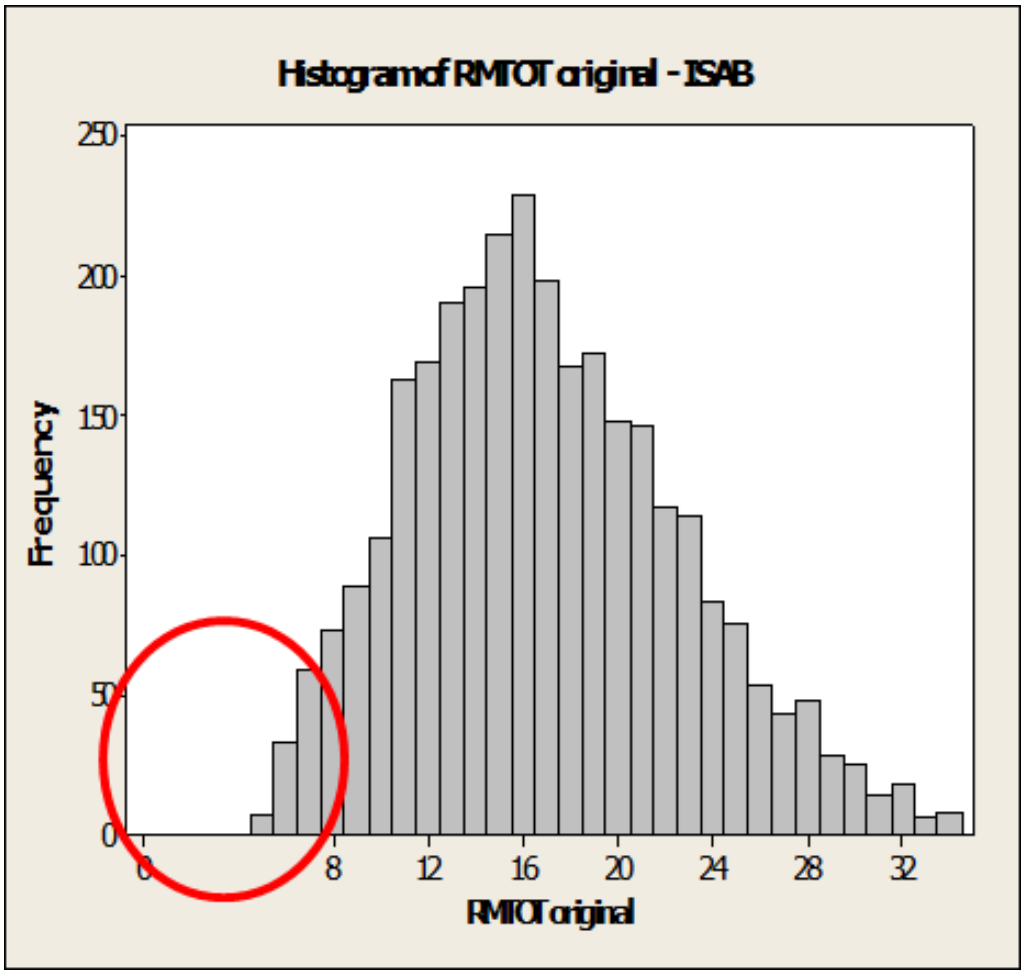


Figura 13 - Histograma do RMTOT do banco de dados original, destacando em vermelho região de falta de dados - domínio ISAB

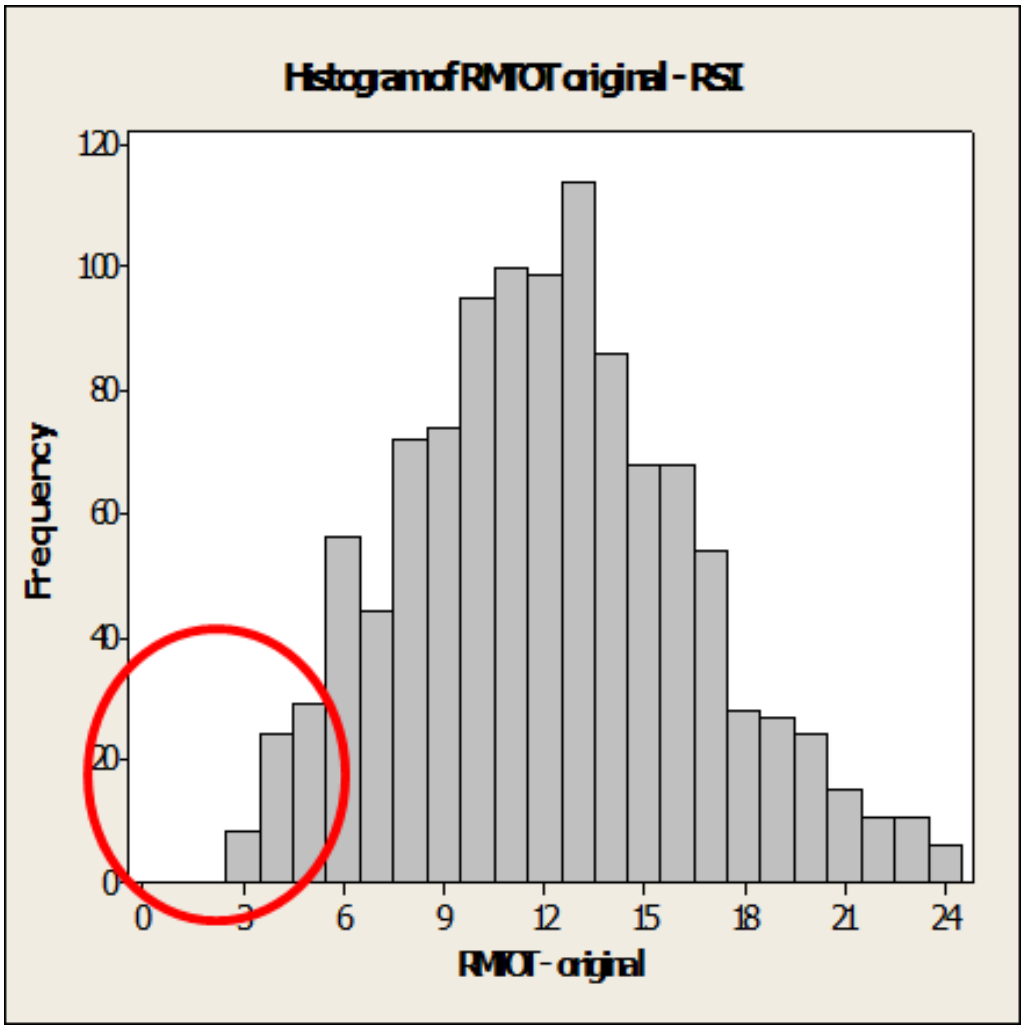


Figura 14 - Histograma do RMTOT do banco de dados original, destacando em vermelho região de falta de dados - domínio RSI

3.1.1 Estatística dos dados por domínio

Para entendimento e interpretação dos dados, foi avaliada a estatística univariada do banco de dados para os dois domínios (Tabelas 2 e 3) e a correlação entre as variáveis químicas e a variável metalúrgica (Tabelas 4 e 5). Essa análise foi realizada apenas para as amostras com resultados completos, pois necessita de isotopia entre os dados.

Tabela 2 - Estatística univariada dos dados originais de ISAB.

Variável	N. amostras	Média	Desvio pad.	Variância	Mínimo	Máximo
Fe_2O_3	3221	27,03	7,96	63,39	9,05	98,72
Al_2O_3	3221	4,41	2,02	4,06	0,01	22,14
MgO	3221	4,29	2,8	7,84	0,05	14,51
SiO_2	3221	22,24	7,4	54,73	1	56,64
CaO	3221	14,58	4,19	17,52	0,81	29,98
TiO_2	3221	9,11	4,42	19,55	1,03	32,69
RCP	3221	1,54	0,49	0,24	0,15	7,18
P_2O_5ap	3221	9,38	2,88	8,27	0,35	21,06
RMTOT	3004	16,94	5,70	32,59	5,29	34,27

Tabela 3 - Estatística univariada dos dados originais de RSI.

Variável	N. amostras	Média	Desvio pad.	Variância	Mínimo	Máximo
Fe_2O_3	1749	19,36	5,29	27,99	7,36	67,96
Al_2O_3	1749	3,35	1,71	2,93	0,01	15,65
MgO	1749	8,92	2,86	8,20	0,66	19,85
SiO_2	1749	27,71	6,21	38,59	3,29	68,29
CaO	1749	17,83	4,26	18,16	1,49	33,1
TiO_2	1749	6,073	3,15	9,95	0,73	34,53
RCP	1749	2,57	0,59	0,36	0,34	3,5
P_2O_5ap	1749	7,18	2,08	4,34	0,53	23,07
RMTOT	1113	12,18	4,31	18,59	3,05	23,98

Tabela 4 - Estatística multivariada dos dados originais de ISAB.

ISAB	Fe_2O_3	Al_2O_3	MgO	SiO_2	CaO	TiO_2	RCP	P_2O_5ap	RMTOT
Fe_2O_3	1								
Al_2O_3	-0,036	1							
MgO	-0,609	0,006	1						
SiO_2	-0,713	0,033	0,581	1					
CaO	-0,302	-0,407	-0,029	-0,211	1				
TiO_2	0,391	-0,188	-0,443	-0,356	-0,291	1			
RCP	-0,526	-0,316	0,653	0,45	0,301	-0,18	1		
P_2O_5ap	-0,021	-0,221	-0,397	-0,438	0,815	-0,217	-0,28	1	
RMTOT	**0**	**-0,161**	**-0,364**	**-0,404**	**0,71**	**-0,181**	**-0,228**	**0,861**	**1**

Tabela 5 - Estatística multivariada dos dados originais de RSI.

RSI	Fe_2O_3	Al_2O_3	MgO	SiO_2	CaO	TiO_2	RCP	P_2O_5ap	RMTOT
Fe_2O_3	1								
Al_2O_3	-0,078	1							
MgO	-0,447	-0,202	1						
SiO_2	-0,425	0,3	0,034	1					
CaO	-0,361	-0,491	0,273	-0,458	1				
TiO_2	0,416	-0,02	-0,456	-0,212	-0,256	1			
RCP	-0,359	-0,224	0,47	0,118	0,375	-0,076	1		
P_2O_5ap	-0,037	-0,229	-0,188	-0,491	0,583	-0,186	-0,502	1	
RMTOT	**-0,088**	**-0,093**	**-0,043**	**-0,205**	**0,414**	**-0,064**	**-0,264**	**0,606**	**1**

Tanto para o domínio ISAB quanto para o RSI, é possível perceber que as variáveis que impactam mais diretamente na recuperação mássica RMTOT são os teores de P_2O_5ap seguido pelos teores de CaO, sendo que estas variáveis também possuem alta correlação entre si. Os demais teores possuem correlação negativa com a variável RMTOT, ou seja, a medida que esses teores crescem a recuperação mássica diminui e essa correlação é menos significativa.

A alta correlação entre as variáveis P_2O_5ap e CaO indica que elas são redundantes, ou seja, a consideração delas para determinar valores de RMTOT pode acarretar em estimativas incorretas. Portanto foi desconsiderado o teor de CaO para as próximas etapas dessa dissertação. As demais variáveis SiO_2, MgO, Fe_2O_3, Al_2O_3, TiO_2 e RCP tem menor influência sobre o rendimento mássico, mas tem sua parcela de contribuição. Contudo, quando as variáveis SiO_2, MgO e RCP são avaliadas conjuntamente sua influência sobre a RMTOT aumenta significativa. Além disso, essas variáveis são importantes de serem controladas na usina de beneficiamento, pois são contaminantes que influenciam na taxa de concentração do produto. Portanto nas etapas seguintes dessa dissertação serão

consideradas as variáveis químicas P_2O_5ap, SiO_2, MgO e RCP e a variável metalúrgica RMTOT.

O objetivo desse trabalho é realizar a modelagem geometalúrgica da variável RMTOT, que é sub amostrada no conjunto de dados. Para isso, serão aplicadas as metodologias propostas nessa dissertação, que consistem no preenchimento das amostras faltantes por atualização bayesiana e aplicação da transformação fixa baseada na distância *inter-quartil* da distribuição do erro para obter o banco de dados complementado. Neste, será aplicada a regressão ACE para gerar do modelo geometalúrgico da mina de fosfato em estudo.

3.2 APLICAÇÃO DA ATUALIZAÇÃO BAYESIANA PARA COMPLEMENTAÇÃO DA VARIÁVEL METALÚRGICA (RMTOT)

A metodologia de atualização bayesiana assume que os dados têm distribuição gaussiana. Os dados não são gaussianos, portando é necessário transformá-los em gaussiano aplicando a rotina estilo Gslib (Deutsch & Journel, 1998) **nscoremv.exe** (Barnett, 2011), nos mesmos. Os resultados dessa transformação são ilustrados nas figuras 15 e 16, sendo RMTOT, SiO_2, P_2O_5AP, MgO, RCP, as variáveis em espaço original e NS: RMTOT, NS:SiO_2, NS:P_2O_5AP, NS:MgO, NS:RCP em espaço gaussiano.

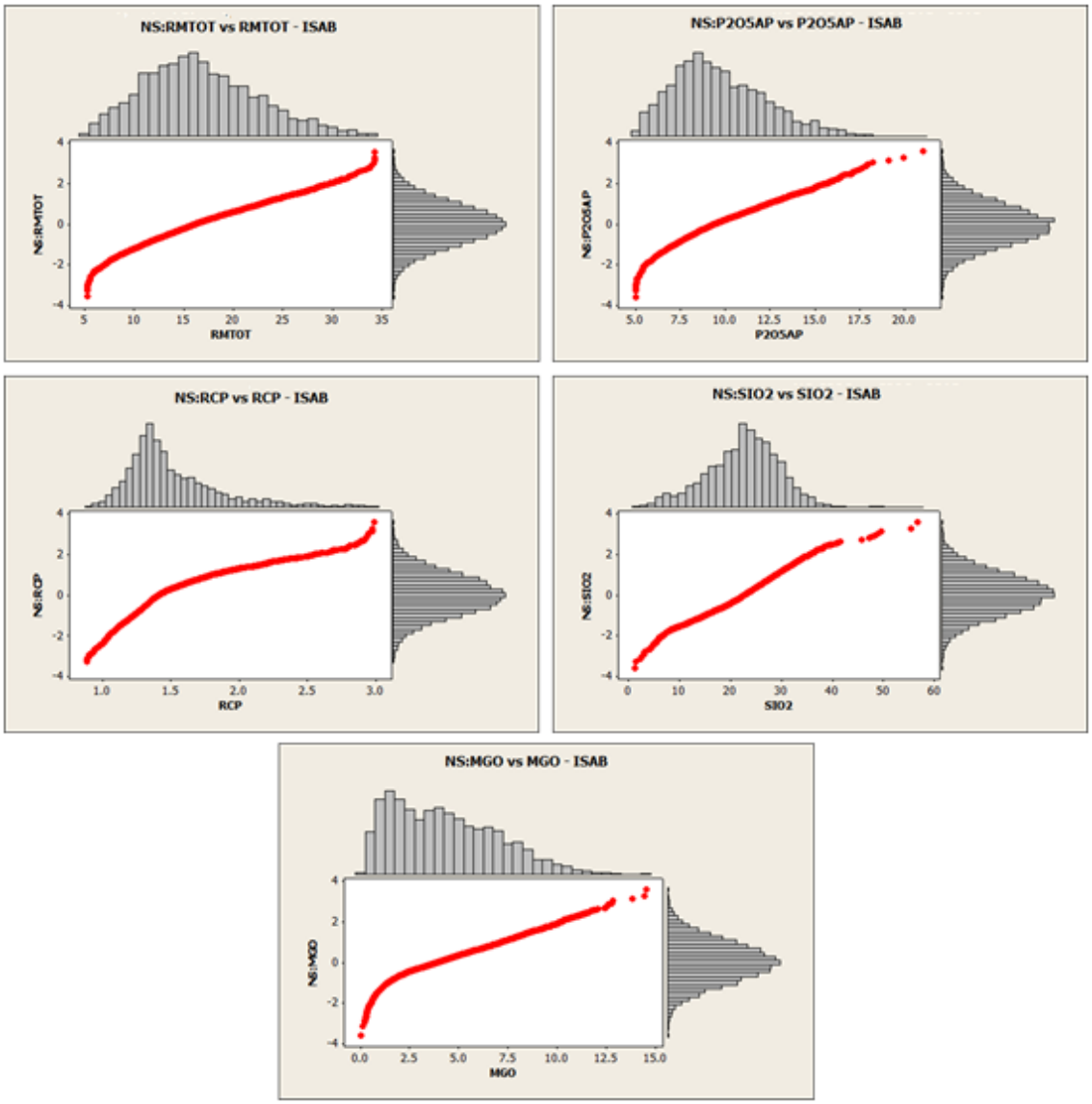


Figura 15 - Transformação das variáveis para espaço gaussiano - domínio ISAB.

Nota-se que no domínio ISAB todas as variáveis antes da transformação têm distribuição assimétrica positiva, sendo a assimetria mais acentuada da variável RCP e menos acentuada da SiO_2 e do RMTOT.

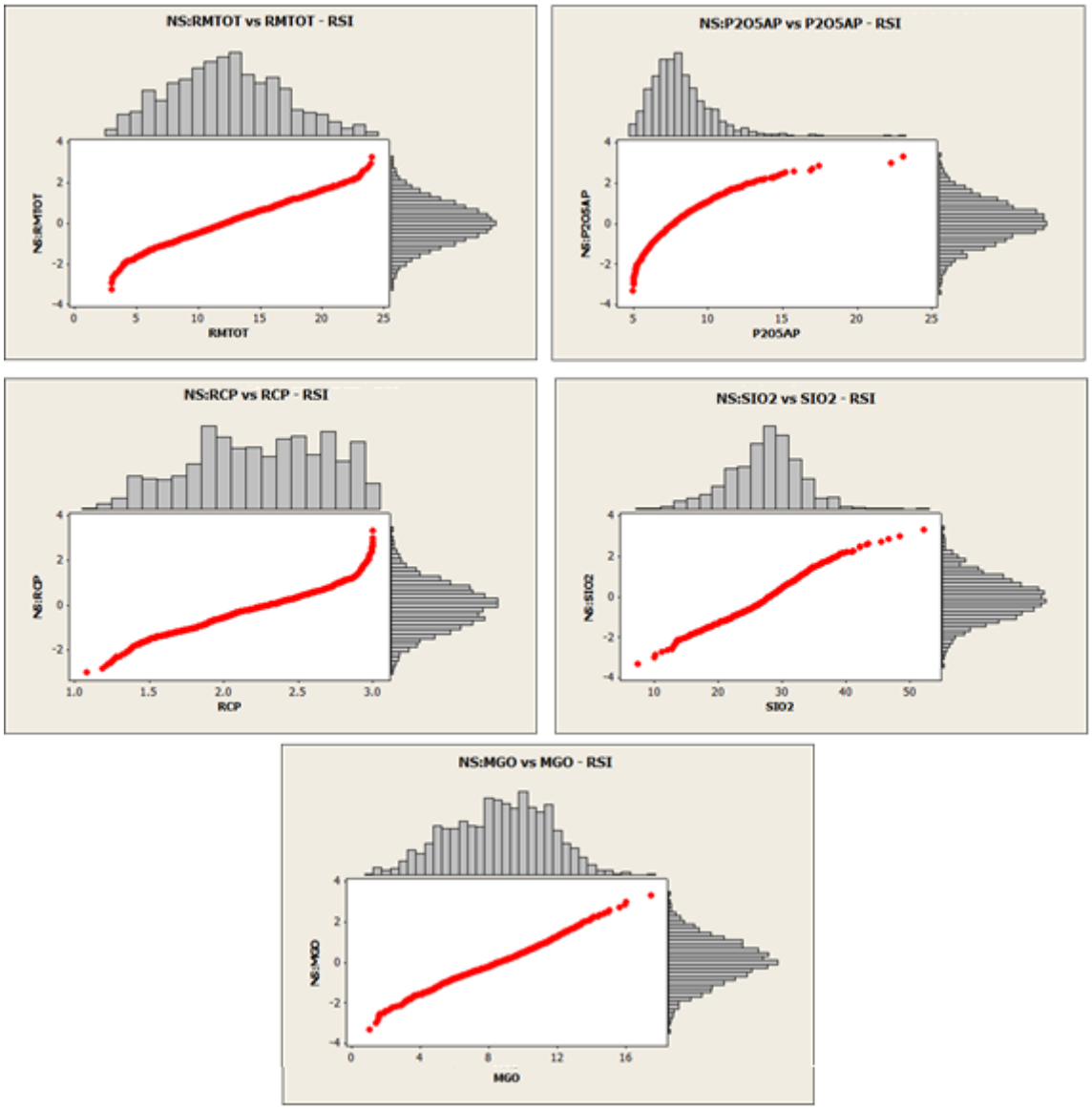


Figura 16 - Transformação das variáveis para espaço gaussiano - domínio RSI.

No domínio RSI, antes da transformação, a variável P_2O_5ap tem distribuição assimétrica positiva, a variável RCP tem distribuição assimétrica negativa e as demais têm distribuição aproximadamente simétrica. Após a transformação as variáveis foram variografadas para serem utilizadas na atualização bayesiana. Os modelos variográficos do domínio ISAB são apresentados nas equações 42 a 56 e do domínio RSI nas equações 47 a 51.

Os variogramas foram ajustados utilizando 20 espaçamentos de 50m, com tolerância de espaçamento de 25metros. A tolerância angular utilizada foi de 22,5° e largura de banda de 25 metros. Os variogramas são direcionais segundo as direções

principais N157, N67, vertical e suas continuidades espaciais são apresentadas nas equações 42 a 51.

Variogramas do domínio ISAB:

- RMTOT

$$\gamma(h) = 0{,}1 + 0{,}55sph \times \left(\frac{N_{157}}{50m}, \frac{N_{67}}{45m}, \frac{D_{90}}{10m}\right) + 0{,}35sph \times \left(\frac{N_{157}}{450m}, \frac{N_{67}}{300m}, \frac{D_{90}}{50m}\right) \quad (42)$$

- P_2O_5ap

$$\gamma(h) = 0{,}1 + 0{,}52sph \times \left(\frac{N_{157}}{45m}, \frac{N_{67}}{40m}, \frac{D_{90}}{5m}\right) + 0{,}38sph \times \left(\frac{N_{157}}{500m}, \frac{N_{67}}{370m}, \frac{D_{90}}{50m}\right) \quad (43)$$

- RCP

$$\gamma(h) = 0{,}1 + 0{,}3sph \times \left(\frac{N_{157}}{50m}, \frac{N_{67}}{45m}, \frac{D_{90}}{10m}\right) + 0{,}6sph \times \left(\frac{N_{157}}{550m}, \frac{N_{67}}{500m}, \frac{D_{90}}{30m}\right) \quad (44)$$

- SiO_2

$$\gamma(h) = 0{,}1 + 0{,}5sph \times \left(\frac{N_{157}}{60m}, \frac{N_{67}}{50m}, \frac{D_{90}}{25m}\right) + 0{,}4sph \times \left(\frac{N_{157}}{400m}, \frac{N_{67}}{300m}, \frac{D_{90}}{30m}\right) \quad (45)$$

- MgO

$$\gamma(h) = 0{,}1 + 0{,}4sph \times \left(\frac{N_{157}}{50m}, \frac{N_{67}}{45m}, \frac{D_{90}}{15m}\right) + 0{,}5sph \times \left(\frac{N_{157}}{350m}, \frac{N_{67}}{200m}, \frac{D_{90}}{25m}\right) \quad (46)$$

Variogramas do domínio RSI:

- RMTOT

$$\gamma(h) = 0{,}1 + 0{,}55sph \times \left(\frac{N_{157}}{50m}, \frac{N_{67}}{45m}, \frac{D_{90}}{10m}\right) + 0{,}35sph \times \left(\frac{N_{157}}{550m}, \frac{N_{67}}{500m}, \frac{D_{90}}{30m}\right) \quad (47)$$

- P_2O_5ap

$$\gamma(h) = 0{,}1 + 0{,}45sph \times \left(\frac{N_{157}}{50m}, \frac{N_{67}}{20m}, \frac{D_{90}}{10m}\right) + 0{,}45sph \times \left(\frac{N_{157}}{320m}, \frac{N_{67}}{220m}, \frac{D_{90}}{45m}\right) \quad (48)$$

- RCP

$$\gamma(h) = 0{,}1 + 0{,}5sph \times \left(\frac{N_{157}}{50m}, \frac{N_{67}}{40m}, \frac{D_{90}}{10m}\right) + 0{,}4sph \times \left(\frac{N_{157}}{500m}, \frac{N_{67}}{300m}, \frac{D_{90}}{50m}\right) \quad (49)$$

- SiO_2

$$\gamma(h) = 0{,}1 + 0{,}45sph \times \left(\frac{N_{157}}{45m}, \frac{N_{67}}{40m}, \frac{D_{90}}{20m}\right) + 0{,}4sph \times \left(\frac{N_{157}}{550m}, \frac{N_{67}}{500m}, \frac{D_{90}}{50m}\right) \quad (50)$$

- MgO

$$\gamma(h) = 0{,}1 + 0.3sph \times \left(\frac{N_{157}}{70m}, \frac{N_{67}}{60m}, \frac{D_{90}}{15m}\right) + 0{,}6sph \times \left(\frac{N_{157}}{900m}, \frac{N_{67}}{650m}, \frac{D_{90}}{80m}\right) \quad (51)$$

3.2.1 Determinação do fator de transformação sob conjunto de calibragem

Uma vez modelado os variogramas das variáveis normalizadas, pode-se utilizar a toina de estilo Gslib (Deutsch & Journel, 1998) **impute.exe** (Barnett, 2013) para realizar

a complementação por metodologia de atualização bayesiana para determinar o fator de transformação.

Para isso, primeiramente foi considerado do banco de dados original dos dois domínios apenas os dados isotópicos, pois esses dados têm estatística relação entre variáveis conhecida. Desse banco de dados isotópico são retiradas aleatoriamente, por sorteio, 20% dos dados de RMTOT, como pode ser visto nas figuras 17 e 18, onde o gráfico em vermelho representa a variável RMTOT isotópica e o gráfico em preto após a exclusão de 20%. Essa remoção deve ser aleatória para não interferir na estatística da amostra. O conjunto de amostras retiradas é chamado de conjunto de calibragem e a comparação do valor de cada uma dessas amostras com seu valor complementado possibilita a avaliação da qualidade do modelo.

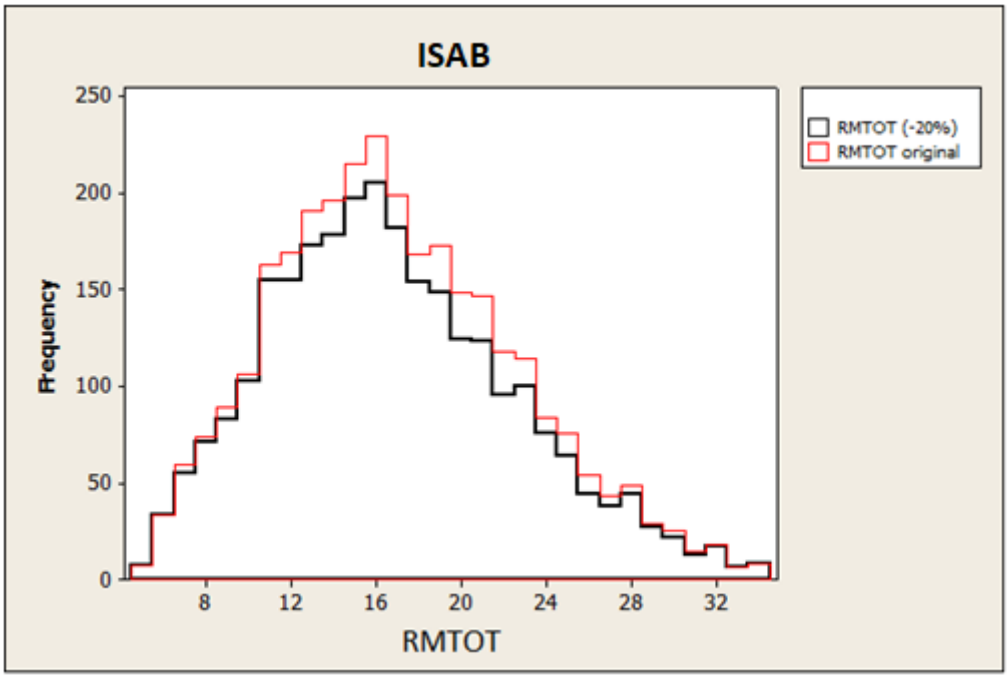


Figura 17 - Histograma RMTOT original isotópico e após remoção de 20% no domínio ISAB.

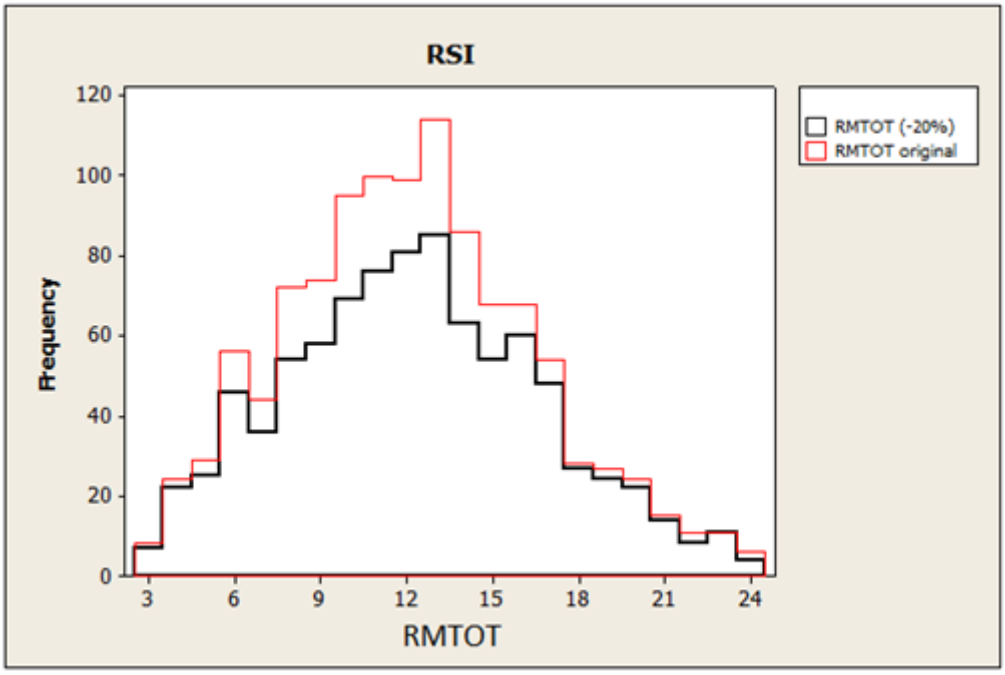


Figura 18 - Histograma RMTOT original isotópico e após remoção de 20% no domínio RSI.

Nota-se que a remoção de 20% dos dados ocorreu de forma aleatório, pois a distribuição dos dados antes e após remoção se manteve a mesma.

A esses conjuntos de calibragem são feitas imputações por atualização bayesiana considerando um máximo de 40 nós previamente simulados, com raios de busca equivalentes aos alcances dos variogramas da variável RMTOT nos dois domínios: 500m na direção N157, 300m na direção N67 e 50m na vertical para o ISAB e 550m na direção N157, 500m na direção N67 e 30m na vertical para o RSI, como pode ser visto nas equações 42 a 51. Cada inserção resultou em 10 cenários completos e manteve-se a estratégia de busca para todas as complementações.

As figuras 19 e 20 apresentam os histogramas acumulados da variável RMTOT do conjunto de calibragem original para os dois domínios ISAB e RSI em vermelho e os 10 cenários complementados em preto.

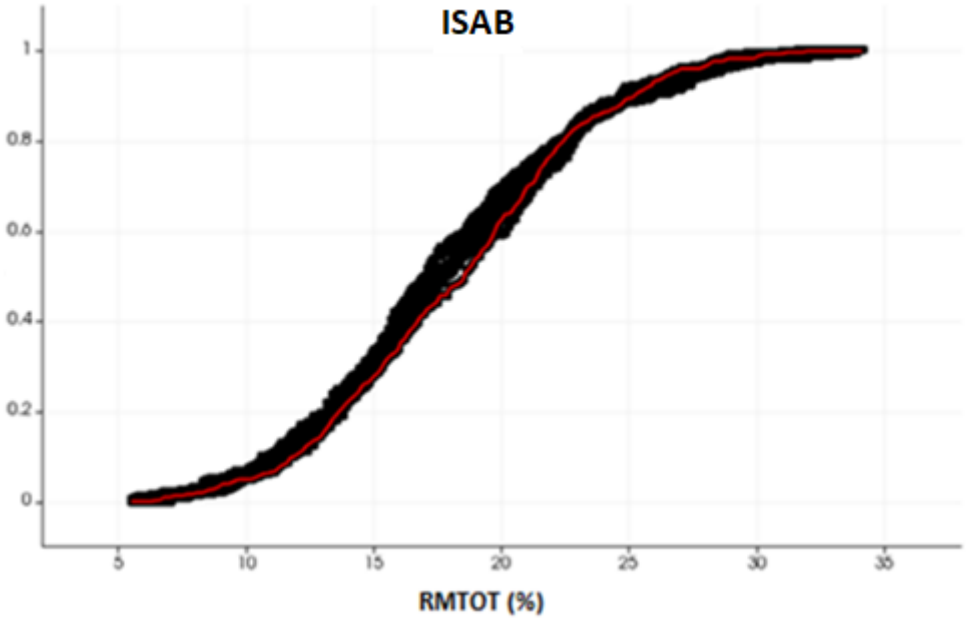


Figura 19 - Histograma acumulado dos dados de RMTOT do conjunto de calibragem original X cenários de imputação do domínio ISAB.

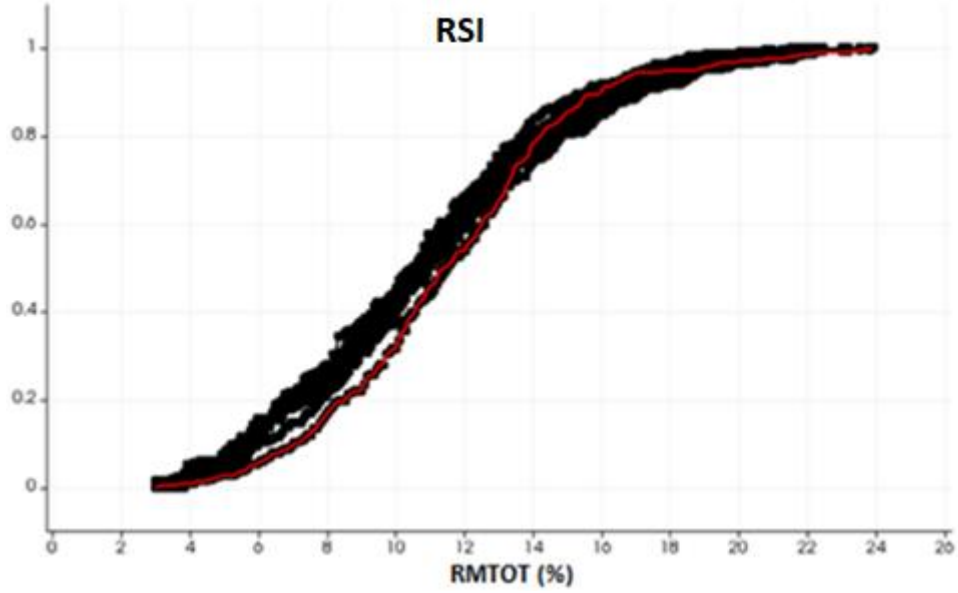


Figura 20 - Histograma acumulado dos dados de RMTOT do conjunto de calibragem original X cenários de imputação do domínio RSI.

A partir das figuras 19 e 20 nota-se que a estimativa do conjunto de calibragem no domínio ISAB foi mais aderente que do domínio RSI no qual ocorreu uma subestimativa dos valores baixos de RMTOT acarretando em médias mais baixas e desvios padrões mais altos nos cenários modelados. Por consequência, tem coeficientes de variação maiores. Para o domínio ISAB, as médias dos 10 cenários variaram pouco em relação à média original, porém os desvios padrões foram maiores acarretando em

maiores coeficientes de variação. Esses vieses observados são considerados aceitáveis, pois é sabido que o mecanismo de falta do banco de dados original é MNAR e a metodologia aplicada para a imputação é adequada para mecanismo de falta MAR, como é descrito no capítulo 2. Assim os modelos foram considerados válidos.

Foi avaliado o erro relativo dos cenários complementados do conjunto de calibragem aplicando a equação 27, amostra a amostra. Os resultados estatísticos dos erros relativos obtidos em cada cenário são apresentados nas tabelas 6 e 7, e a representação gráfica nas figuras 21 e 22.

Tabela 6 - Estatística da distribuição dos erros relativos para os 10 cenários de imputação do conjunto de calibragem do domínio ISAB.

Domínio	Variável	Média	Desvio pad.	Mínimo	Q1	Mediana	Q3	Máximo	IQR
ISAB	erroj1	-0,0062	0,2788	-2,2494	-0,1357	0,0319	0,1794	0,5184	0,3151
	erroj2	-0,0252	0,2471	-1,2183	-0,1337	-0,0184	0,1489	0,5453	0,2826
	erroj3	-0,0205	0,2977	-2,4109	-0,1471	0,0276	0,1514	0,6182	0,2984
	erroj4	-0,0008	0,281	-1,6626	-0,1137	0,0311	0,1827	0,5363	0,2964
	erroj5	-0,0298	0,2917	-2,2163	-0,1565	-0,0053	0,1472	0,5637	0,3037
	erroj6	-0,0108	0,2908	-1,9754	-0,1481	0,0376	0,1875	0,5307	0,335
	erroj7	0,003	0,2612	-1,3207	-0,1178	0,0337	0,1695	0,5495	0,2873
	erroj8	0,0047	0,232	-0,9117	-0,1182	0,0426	0,1737	0,598	0,2919
	erroj9	-0,0125	0,2931	-1,8652	-0,1422	0,0264	0,1763	0,5907	0,3184
	erroj10	0,0007	0,2625	-1,4173	-0,1307	0,0301	0,1796	0,4789	0,3103

Tabela 7 - Estatística da distribuição dos erros relativos para os 10 cenários de imputação do conjunto de calibragem do domínio RSI.

Domínio	Variável	Média	Desvio pad.	Mínimo	Q1	Mediana	Q3	Máximo	IQR
RSI	erroj1	0,009	0,4438	-1,8699	-0,1901	0,0833	0,3182	0,7037	0,5084
	erroj2	-0,0575	0,508	-2,6703	-0,2469	0,0254	0,2768	0,6951	0,5237
	erroj3	-0,0224	0,4726	-2,5931	-0,2343	0,0562	0,2848	0,7441	0,5191
	erroj4	-0,0664	0,5386	-3,0667	-0,3059	0,0326	0,2879	0,8011	0,5937
	erroj5	0,0197	0,4329	-2,3165	-0,1897	0,1038	0,332	0,7127	0,5217
	erroj6	-0,0401	0,4626	-1,9577	-0,237	0,0441	0,2577	0,7722	0,4947
	erroj7	-0,0541	0,6019	-3,8325	-0,2266	0,1111	0,2782	0,764	0,5048
	erroj8	-0,0088	0,5201	-3,8667	-0,1878	0,0978	0,2965	0,695	0,4843
	erroj9	0,0092	0,4397	-1,9542	-0,2071	0,0866	0,3131	0,7359	0,5201
	erroj10	0,0136	0,4705	-2,0401	-0,1964	0,129	0,3165	0,7841	0,5129

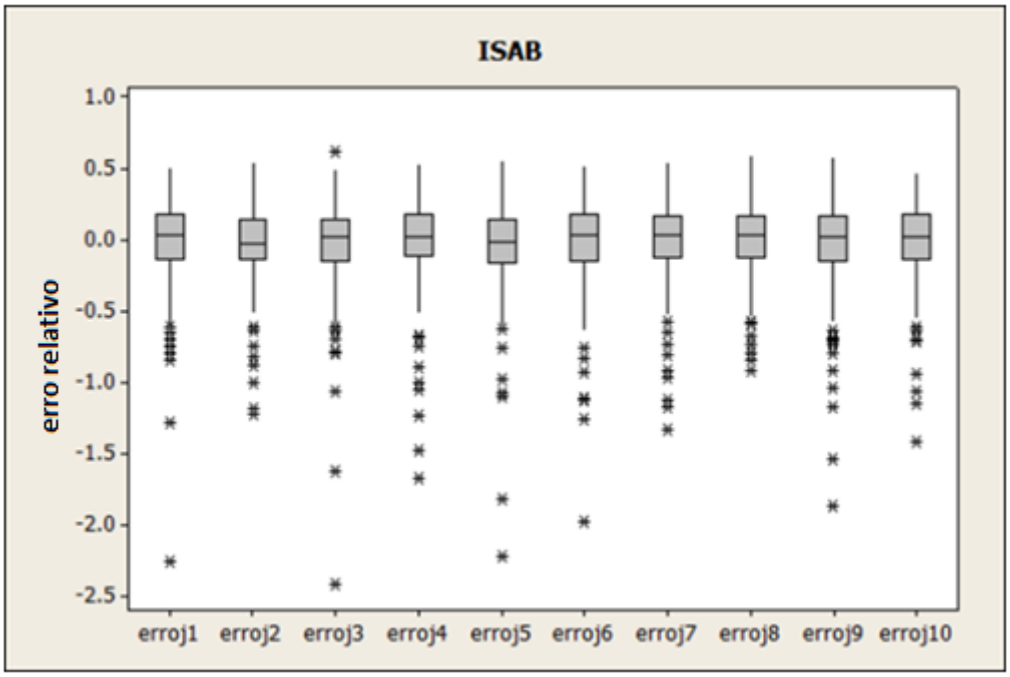


Figura 21 - Gráfico da distribuição do erro relativo para os 10 cenários do domínio ISAB.

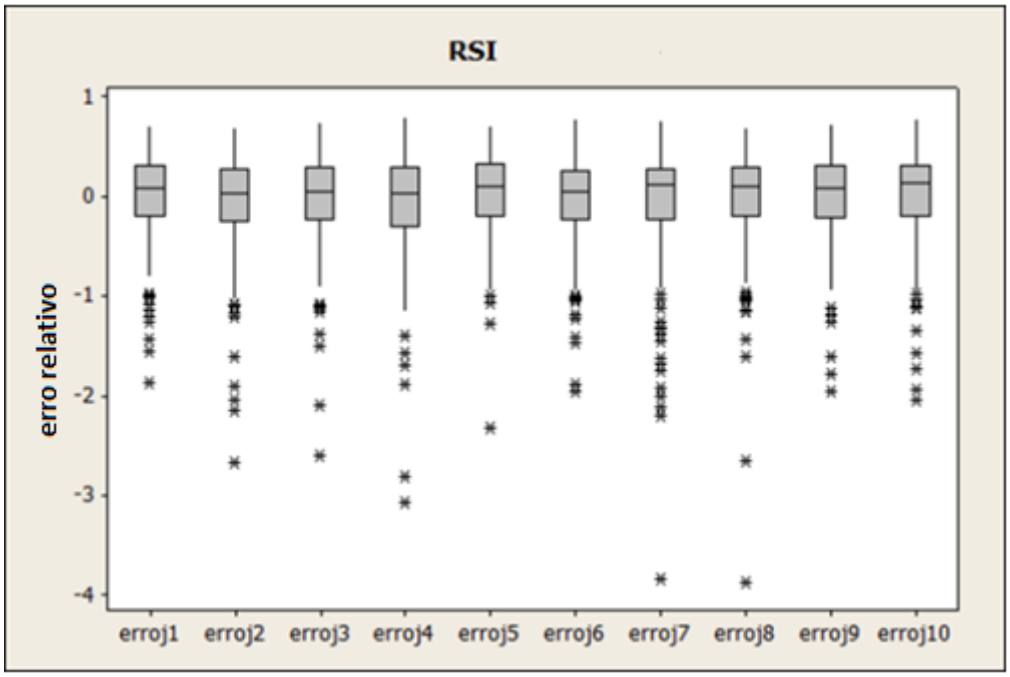


Figura 22 - Gráfico da distribuição do erro relativo para os 10 cenários do domínio RSI.

Nota-se, que o erro relativo médio de cada cenário é próximo de zero, o erro relativo máximo dos cenários do domínio ISAB foi de 240% e do RSI de 386%, o que torna inviável a utilização dessa magnitude de correção como fator de transformação. De acordo com Silva (2018), a utilização do erro relativo máximo como fator de

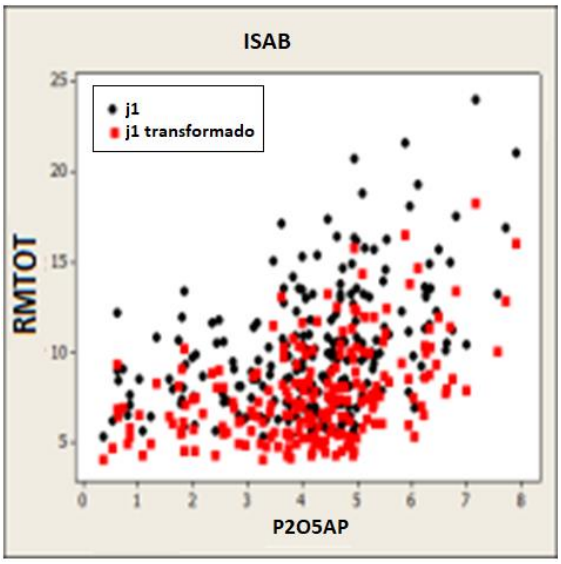
transformação não é recomendada para os dados em questão, pois a distribuição dos mesmos não tem forte assimetria e a região faltante não é no mesmo sentido da assimetria, como se pode ver nas figuras 13 e 14. Para esse caso, em que a variável é aproximadamente simétrica e a falta de dados se dá na região oposta a assimetria, Silva (2018) recomenda a aplicação da transformação fixa proposta por Rubin (1987).

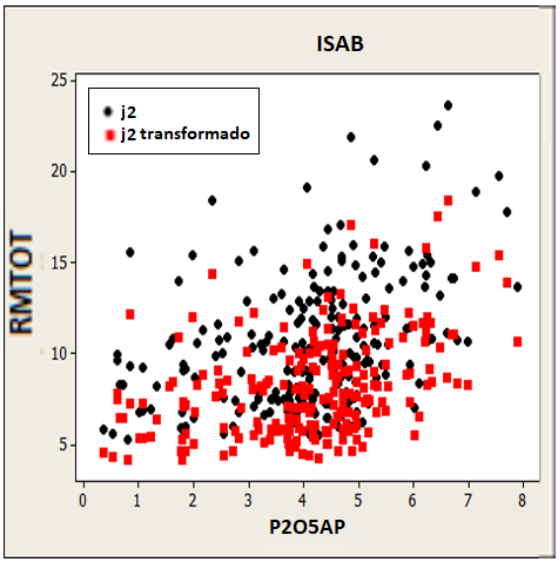
Porém, pela estatística e distribuição dos erros relativos, nota-se que para o domínio ISAB o desvio padrão e distância *inter-quartil* é da ordem de 30% e para o RSI é da ordem de 50%, o que indica que a aplicação da transformação proposta por Rubin (1987) de 20% não obteria resultado satisfatório, por não representar espalhamento do erro nos dois domínios. Desta forma, optou-se como fator de transformação a distância *inter-quartil* (IQR) de cada cenário. Ou seja, das amostras a serem complementadas nos 10 cenários de imputação dos domínios ISAB e RSI serão subtraídos os valores de IQR de seu respectivo valor em cada cenário.

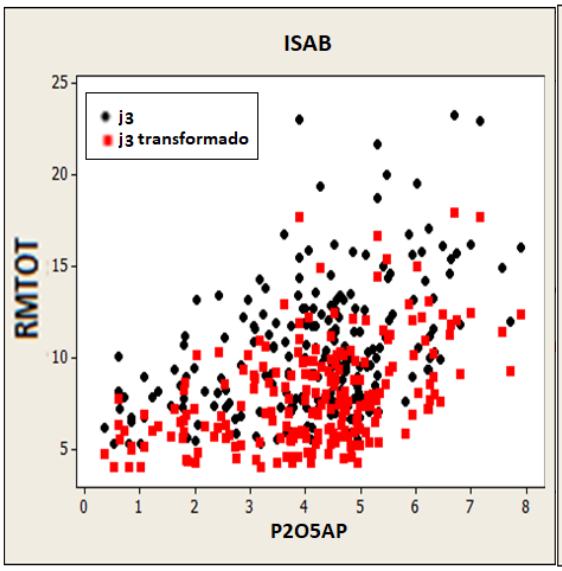
3.2.2 Inserção no banco de dados original MNAR dos domínios ISAB e RSI

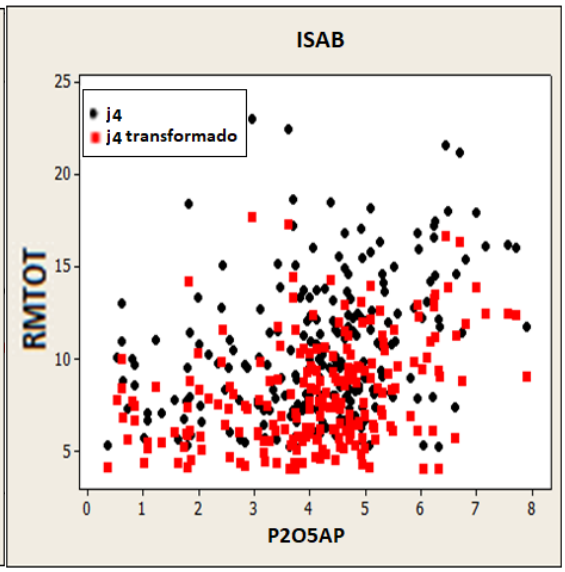
A complementação dos dados faltantes de RMTOT do banco de dados original, assim como no conjunto de calibragem, é realizada por atualização bayesiana (Ren, 2007) utilizando o programa estilo Gslib (Deutsch & Journel, 1998) **impute.exe** (Barnett, 2013). Para tal, são considerados os mesmos parâmetros do conjunto de calibragem: um máximo de 40 nós previamente simulados, com raios de busca equivalentes aos alcances dos variogramas da variável RMTOT nos dois domínios: 500m na direção N157, 300m na direção N67 e 50m na vertical para o ISAB e 550m na direção N157, 500m na direção N67 e 30m na vertical para o RSI. O resultado da inserção são 10 cenários completos para cada domínio.

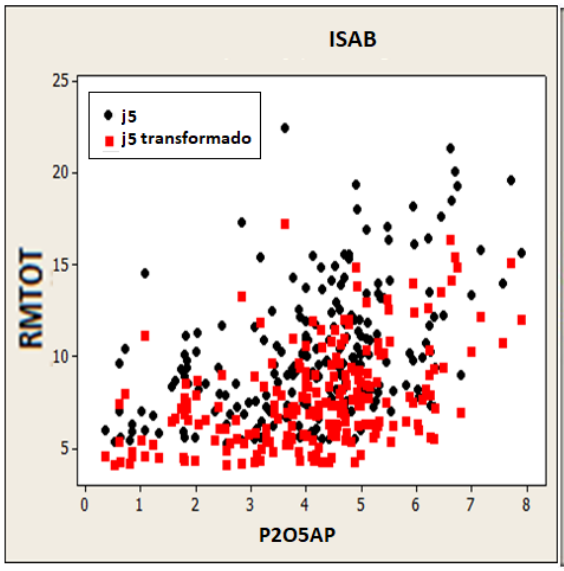
Aos 10 cenários resultantes da atualização bayesiana para os domínios ISAB e RSI, foi aplicada a transformação fixa subtraindo de cada amostra complementada o valor da IQR do erro relativo obtido no respectivo cenário do conjunto de calibragem, como exemplificado nas equações 27 e 28. Gerou-se os 10 cenários com a variável RMTOT completa adequados ao mecanismo de falta MNAR. As figuras 23 e 24 mostram, no gráfico de dispersão de RMTOT por P2O5ap, essa transformação. O deslocamento para baixo no eixo Y entre os pontos pretos (antes da aplicação da transformação) e vermelhos (após transformação) mostra que os valores de RMTOT, nas amostras complementadas, após a transformação fixa são menores.

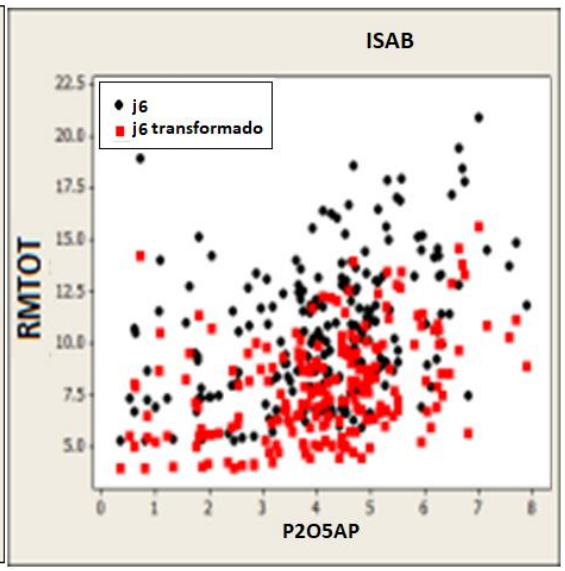

ISAB
j1
j1 transformado
RMTOT
P2O5AP


ISAB
j2
j2 transformado
RMTOT
P2O5AP


ISAB
j3
j3 transformado
RMTOT
P2O5AP


ISAB
j4
j4 transformado
RMTOT
P2O5AP


ISAB
j5
j5 transformado
RMTOT
P2O5AP


ISAB
j6
j6 transformado
RMTOT
P2O5AP

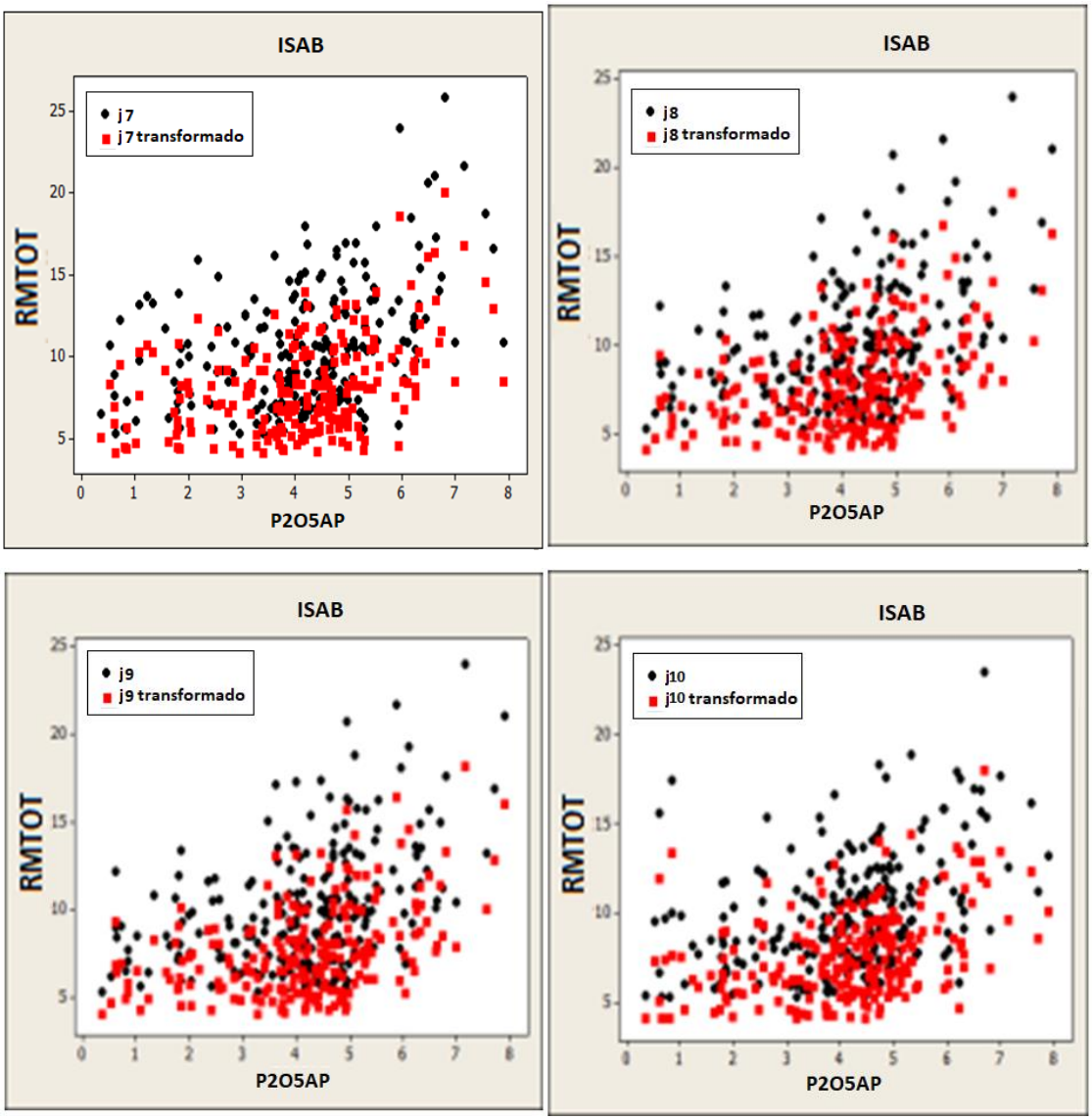
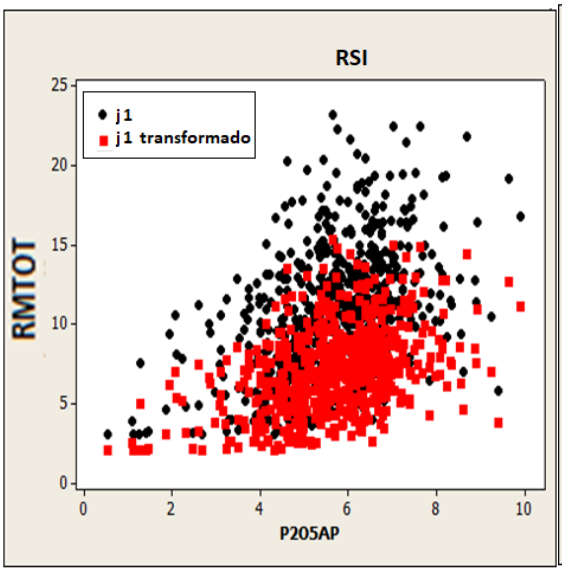
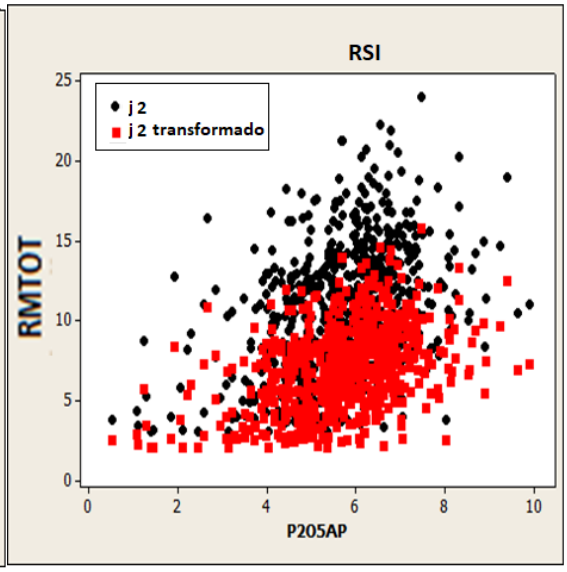
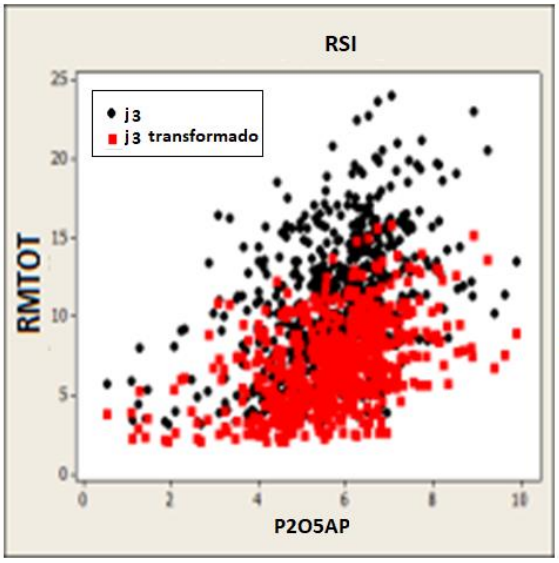
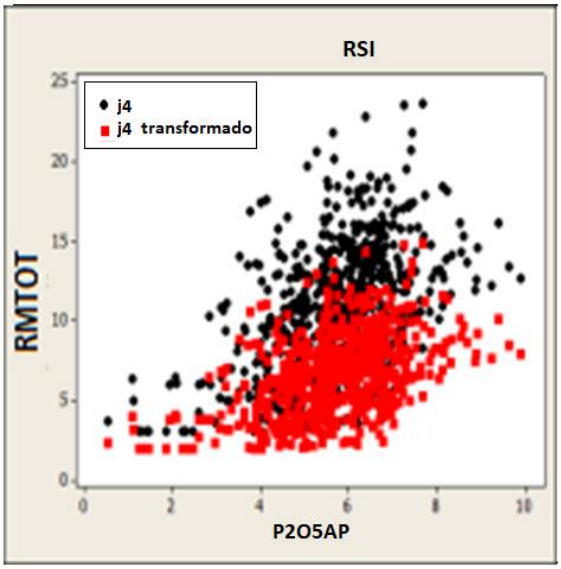


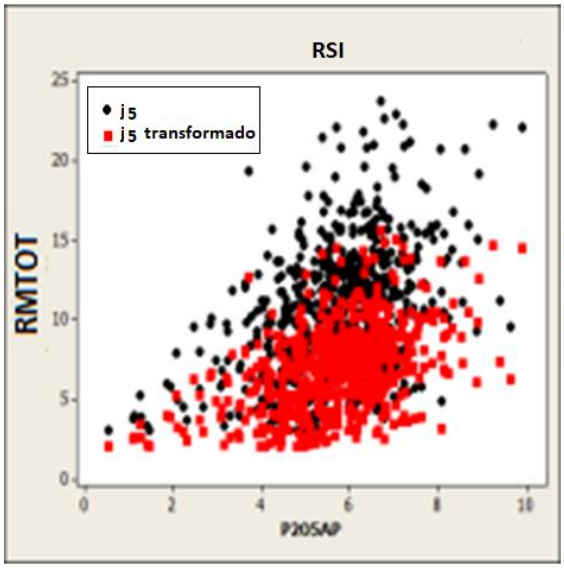
Figura 23 - Gráficos de dispersão do RMTOT x P_2O_5ap das amostras complementadas nos 10 cenários antes e após transformação – ISAB.

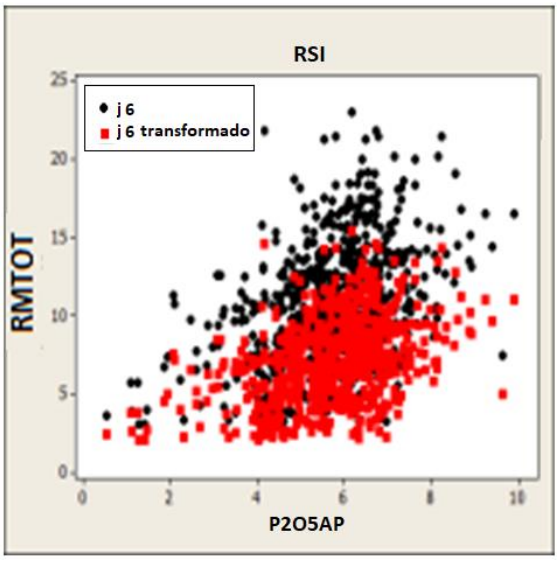

RSI
j 1
j 1 transformado
RMTOT
P205AP


RSI
j 2
j 2 transformado
RMTOT
P205AP


RSI
j 3
j 3 transformado
RMTOT
P2O5AP


RSI
j4
j4 transformado
RMTOT
P2O5AP


RSI
j 5
j 5 transformado
RMTOT
P205AP


RSI
j 6
j 6 transformado
RMTOT
P2O5AP

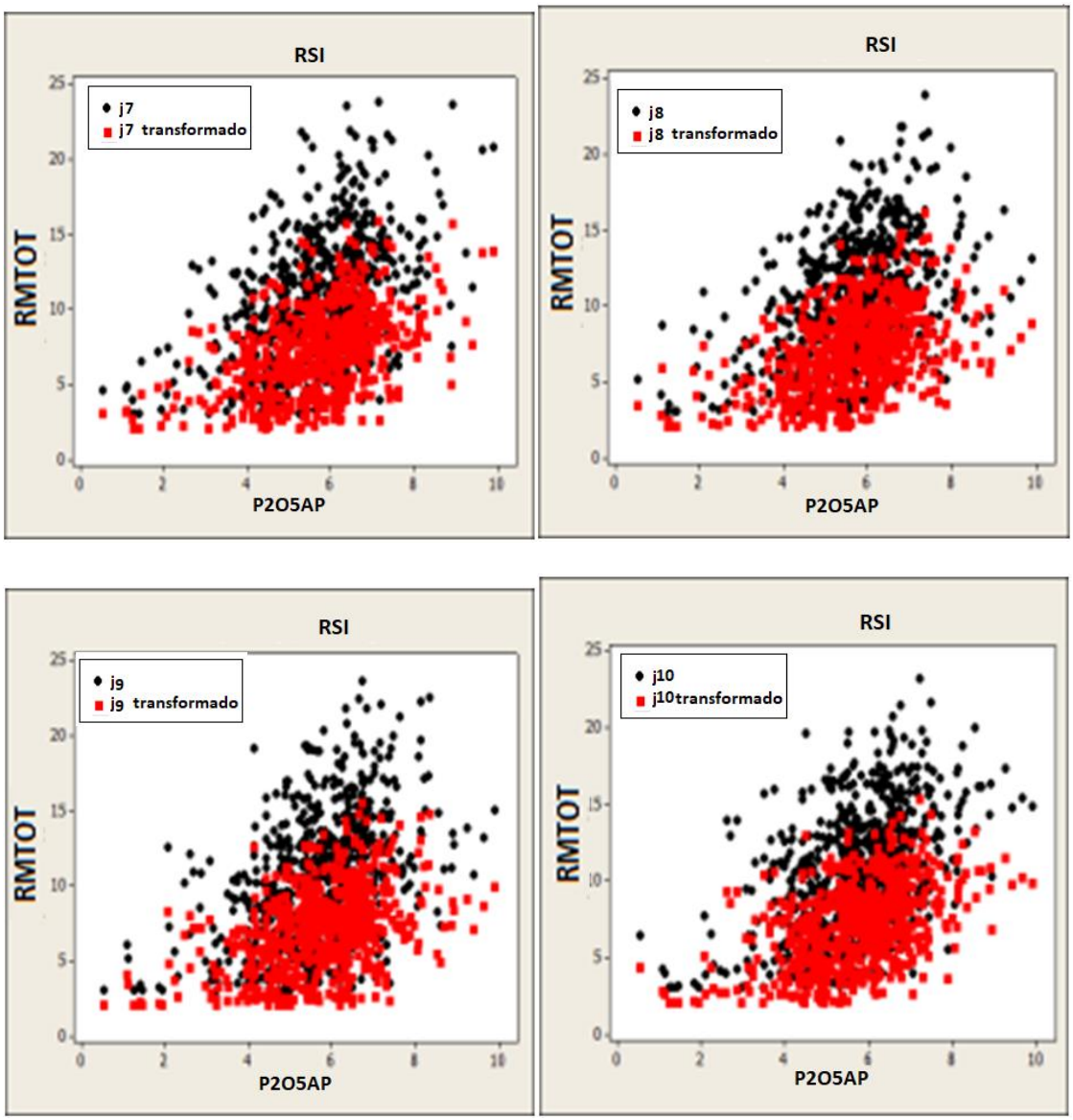


Figura 24 - Gráficos de dispersão do RMTOT x P_2O_5ap das amostras complementadas nos 10 cenários antes e após transformação – RSI.

Ao fim da atualização bayesiana com a transformação fixa obteve-se 10 cenários de banco de dados completos para cada domínio. As figuras 25 e 26 apresentam os histogramas do RMTOT original (linha preta) e dos 10 cenários completos (demais linhas).

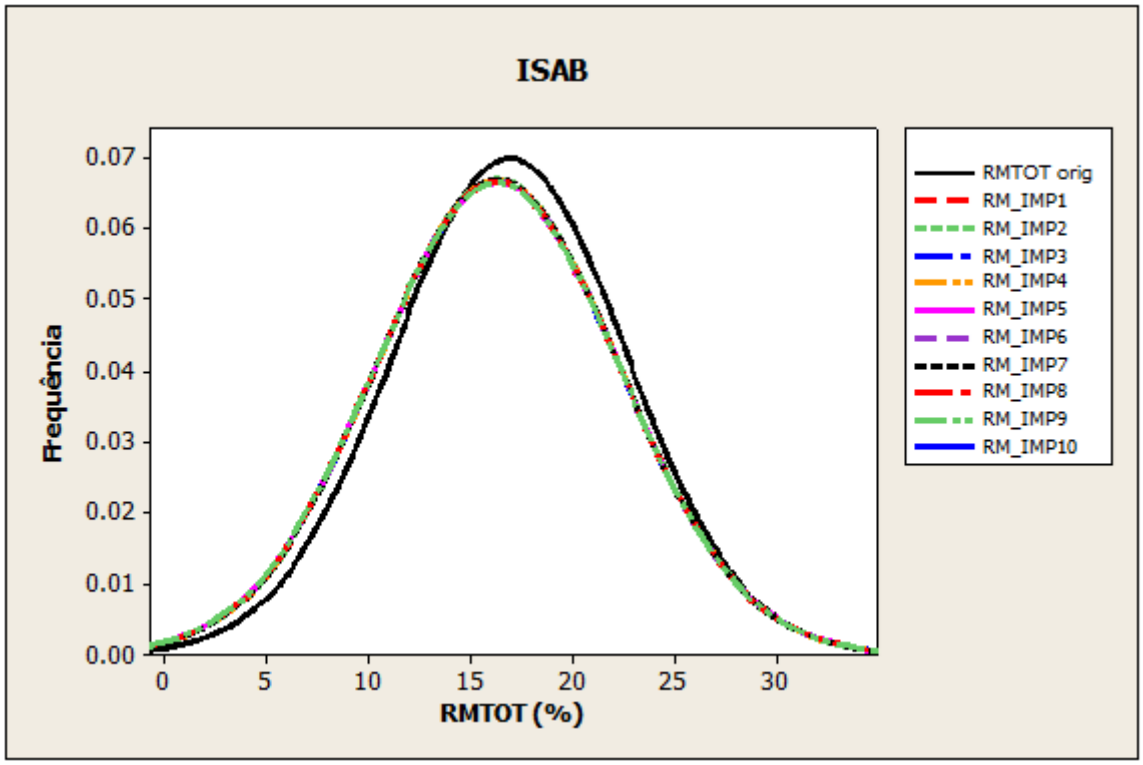


Figura 25 - Histogramas do RMTOT do banco original e dos cenários finais completos - ISAB.

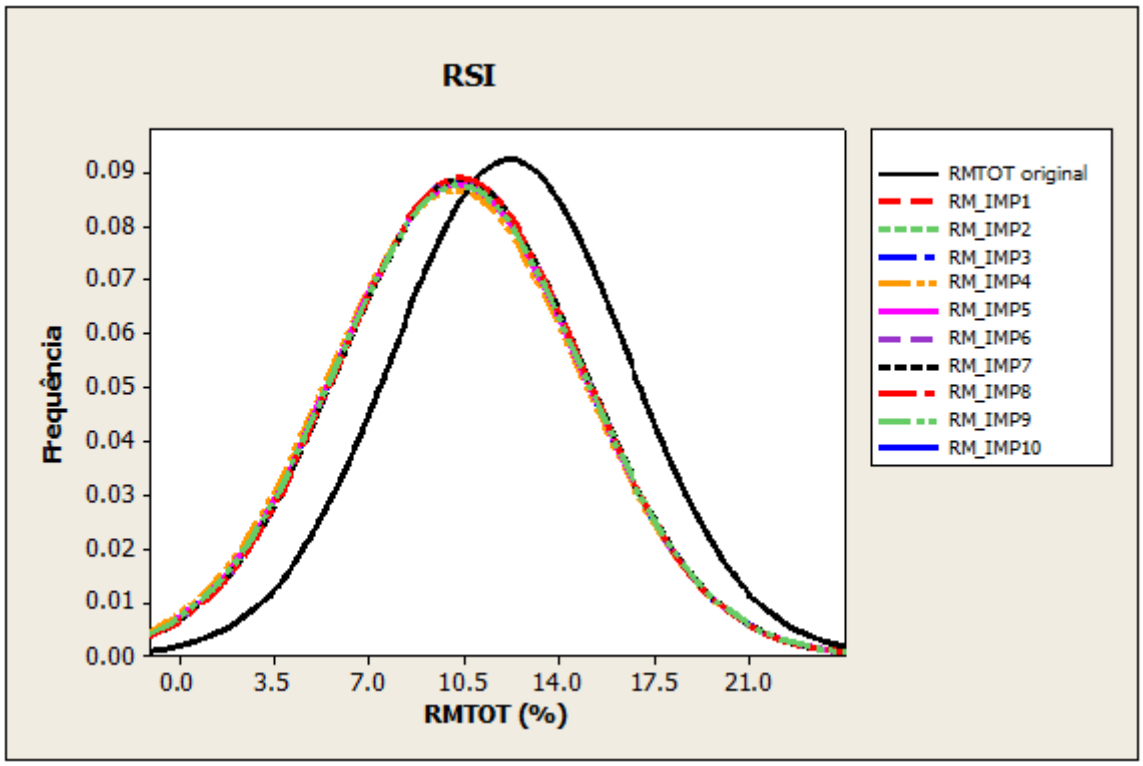


Figura 26 - Histogramas do RMTOT do banco original e dos cenários finais completos - RSI.

Nota-se, que a complementação ocorreu na cauda inferior dos histogramas tornando as médias dos dados menores. A tabela 8 apresenta a estatística completa dos cenários, deixando evidente que a média dos dados originais é maior que dos dados complementados, o valor mínimo nos 10 cenários é menor e o desvio padrão maior, acarretando em um maior coeficiente de variação. A complementação dos dados foi satisfatória ajustando sua distribuição nos baixos valores.

Tabela 8 - Estatística do RMTOT dos dados originais e dos 10 cenários complementados.

Domínio	Cenário	Nº de amostras	Média	Mínimo	Máximo	Desvio pad.	Variância	CV
ISAB	RMTOT orig	3004	16,943	5,29	34,27	5,709	32,59	33,69
	RM_IMP1	3221	16,319	4,023	34,27	6,021	36,251	36,89
	RM_IMP2	3221	16,356	4,124	34,27	5,975	35,696	36,53
	RM_IMP3	3221	16,336	4,074	34,27	6,001	36,011	36,73
	RM_IMP4	3221	16,35	4,081	34,27	5,985	35,815	36,6
	RM_IMP5	3221	16,321	4,06	34,27	6,021	36,247	36,89
	RM_IMP6	3221	16,331	3,963	34,27	6.00	35,997	36,74
	RM_IMP7	3221	16,35	4,109	34,27	5,985	35,826	36,61
	RM_IMP8	3221	16,328	4,095	34,27	6,01	36,115	36,81
	RM_IMP9	3221	16,319	4,012	34,27	6,022	36,267	36,9
	RM_IMP10	3221	16,317	4,037	34,27	6,02	36,242	36,9
RSI	RMTOT orig	1113	12,182	3,05	23,98	4,312	18,597	35,4
	RM_IMP1	1749	10,338	2,022	23,98	4,507	20,315	43,6
	RM_IMP2	1749	10,308	2,002	23,98	4,526	20,481	43,9
	RM_IMP3	1749	10,292	2,008	23,98	4,563	20,822	44,34
	RM_IMP4	1749	10,174	1,914	23,98	4,611	21,265	45,32
	RM_IMP5	1749	10,285	2,004	23,98	4,557	20,764	44,31
	RM_IMP6	1749	10,361	2,041	23,98	4,5	20,252	43,44
	RM_IMP7	1749	10,382	2,027	23,98	4,509	20,329	43,43
	RM_IMP8	1749	10,362	2,055	23,98	4,495	20,204	43,38
	RM_IMP9	1749	10,301	2,006	23,98	4,564	20,83	44,31
	RM_IMP10	1749	10,31	2,02	23,98	4,54	20,613	44,04

3.3 APLICAÇÃO DA ESPERANÇA CONDICIONAL ALTERNADA (ACE) PARA ESTIMATIVA DO MODELO GEOMETALÚRGICO

Os 10 cenários complementados, de cada domínio, obtidos na etapa anterior serão utilizados como base de dados para gerar o modelo geometalúrgico do depósito em questão por ACE. Para isso, cada cenário do domínio ISAB será agrupado ao seu respectivo cenário do domínio RSI, de forma a obter 10 cenários do banco de dados total. Esse agrupamento é realizado nessa etapa, pois na operação de mina os domínios ISAB e RSI não são individualizados, impossibilitando o rastreamento deles na alimentação da usina de beneficiamento. Além desses 10, também é utilizado um 11° cenário, que é a

média entre os 10 obtidos anteriormente, chamado nessa etapa do trabalho de cenário médio (jmédio). O banco de dados original com amostras incompletas, chamado de cenário faltante (jfalt) também é considerado para gerar um modelo geometalúrgico, assim poder-se-á avaliar se houve ganho de acuracidade e precisão aos modelos de regressão gerados após completar-se os dados.

A primeira abordagem de um modelo de regressão é verificar a relação entre as variáveis preditoras (teores) e a variável resposta (RMTOT) pelo o coeficiente de correlação. As tabelas9 e 10 apresentam as matrizes de correlação entre essas variáveis para o banco de dados original faltante e o banco de dados complementado com valores do cenário médio.

Tabela 9 - Matriz de correlação do cenário dos dados originais faltantes.

	Fe_2O_3	Al_2O_3	MgO	SiO_2	CaO	TiO_2	RCP	P_2O_5ap	RMTOT orig
Fe_2O_3	1								
Al_2O_3	0,054	1							
MgO	-0,682	-0,165	1						
SiO_2	-0,721	-0,015	0,531	1					
CaO	-0,405	-0,455	0,259	-0,098	1				
TiO_2	0,474	-0,108	-0,54	-0,434	-0,319	1			
RCP	-0,612	-0,365	0,726	0,462	0,474	-0,313	1		
P_2O_5ap	0,136	-0,111	-0,423	-0,483	0,556	-0,059	-0,439	1	
RMTOT orig	0,168	-0,021	-0,404	-0,424	0,408	0,006	-0,401	0,815	1

Tabela 10 - Matriz de correlação do cenário médio dos dados complementados.

	Fe_2O_3	Al_2O_3	MgO	SiO_2	CaO	TiO_2	RCP	P_2O_5ap	RM_jmédio
Fe_2O_3	1								
Al_2O_3	0,075	1							
MgO	-0,659	-0,226	1						
SiO_2	-0,684	0	0,494	1					
CaO	-0,502	-0,483	0,339	-0,051	1				
TiO_2	0,453	-0,055	-0,551	-0,42	-0,341	1			
RCP	-0,571	-0,381	0,721	0,407	0,5	-0,323	1		
P_2O_5ap	0,062	-0,092	-0,414	-0,421	0,479	-0,034	-0,473	1	
RM_médio	0,161	0,01	-0,452	-0,405	0,299	0,05	-0,505	0,852	1

Nota-se, que a variável que mais impacta na recuperação mássica é o P_2O_5ap, com correlação da ordem de 0,8, tanto no banco de dados original quanto no banco de dados do cenário médio. Em seguida são: RCP, MgO, SiO_2 e CaO. Os óxidos Fe_2O_3, TiO_2 e Al_2O_3 tem correlações muito baixas com o RMTOT e não são estatisticamente

significativos para o modelo em questão. O óxido CaO, apesar de ter uma correlação de 0,4 com o RMTOT no cenário de dados originais e 0,299 no cenário de dados complementados médio, não foi utilizado na regressão. P_2O_5ap e CaO são redundantes por terem alta correlação entre si, ou seja, a consideração dessas variáveis para determinar valores de RMTOT pode acarretar em estimativas incorretas. O RCP apesar de ter correlação com o P_2O_5ap, por ser uma variável derivada do CaO e P_2O_5, ela representa a taxa de carbonato que são minerais contaminantes que influenciam no potencial de recuperação da apatita. O MgO e SiO_2 são variáveis que indicam a porcentagem de minerais silicatos que também influenciam na flotação. Assim, para a modelagem ACE foram consideradas as variáveis preditoras P_2O_5ap, RCP, MgO e SiO_2 e a variável resposta o RMTOT.

A regressão ACE é aplicada nos dados em questão, gerando uma função de transformação θ_RMTOT da variável resposta RMTOT, e funções de transformações $\varphi_{P2O5ap}, \varphi_{RCP}, \varphi_{MgO}\ e\ \varphi_{SiO2}$ das respectivas variáveis preditoras P_2O_5ap, RCP, MgO e SiO_2. Essas funções transformações são geradas para todos os cenários listados: o cenário original com dados faltantes, o cenário médio complementado e os 10 cenários resultantes da imputação. Essas funções transformações são estimadas de forma iterativa pela rotina estilo Gslib (Deutsch & Journel, 1998) **ace.exe** (Barnett, 2013), buscando minimizar a variância inexplicada da relação linear entre a variável resposta transformada e a soma das variáveis preditoras transformadas. Como as funções transformação consistem em um valor de transformação para cada amostra é inviável apresentá-las em forma de tabela, assim as mesmas são apresentadas pelas figuras 28 e 29 em forma de gráficos de dispersão das variáveis versus suas respectivas funções transformações e da função transformação θ_RMTOT versus a soma das funções φ_i, dos dados originais e do cenário médio. Os gráficos dos 10 cenários encontram-se no anexo 1. A figura 27 apresenta o arquivo de parâmetro do **ace.exe** para gerar as funções transformações do cenário médio, para os demais cenários foram alterados os arquivos de entra e saída.

```
                    Parameters for ACE
                    ******************

START OF PARAMETERS:
data_jmedia.txt                          -file with data
5 0                                      - col for response variable and weight
4                                        - # of predictor variables
1 2 3 4                    - cols for predictor variables
-1.0e21     1.0e21                       -trimming limits
1 1 1 1 3                     -variable flag (see NOTE at bottom)
ace_jmedia.out                                -output file for the optimal transforms


NOTE: Variable flag applies as follows:
  l(p+1) : flag for each variable.
  l(1) through l(p) : predictor variables.
  l(p+1) : response variable.
  l(i)=0 => ith variable not to be used.
  l(i)=1 => ith variable assumes orderable values.
  l(i)=2 => ith variable assumes circular (periodic) values
            in the range (0.0,1.0) with period 1.0.
  l(i)=3 => ith variable transformation is to be monotone.
  l(i)=4 => ith variable transformation is to be linear.
  l(i)=5 => ith variable assumes categorical values.
```

Figura 27 - Arquivo de parâmetros do ace.exe

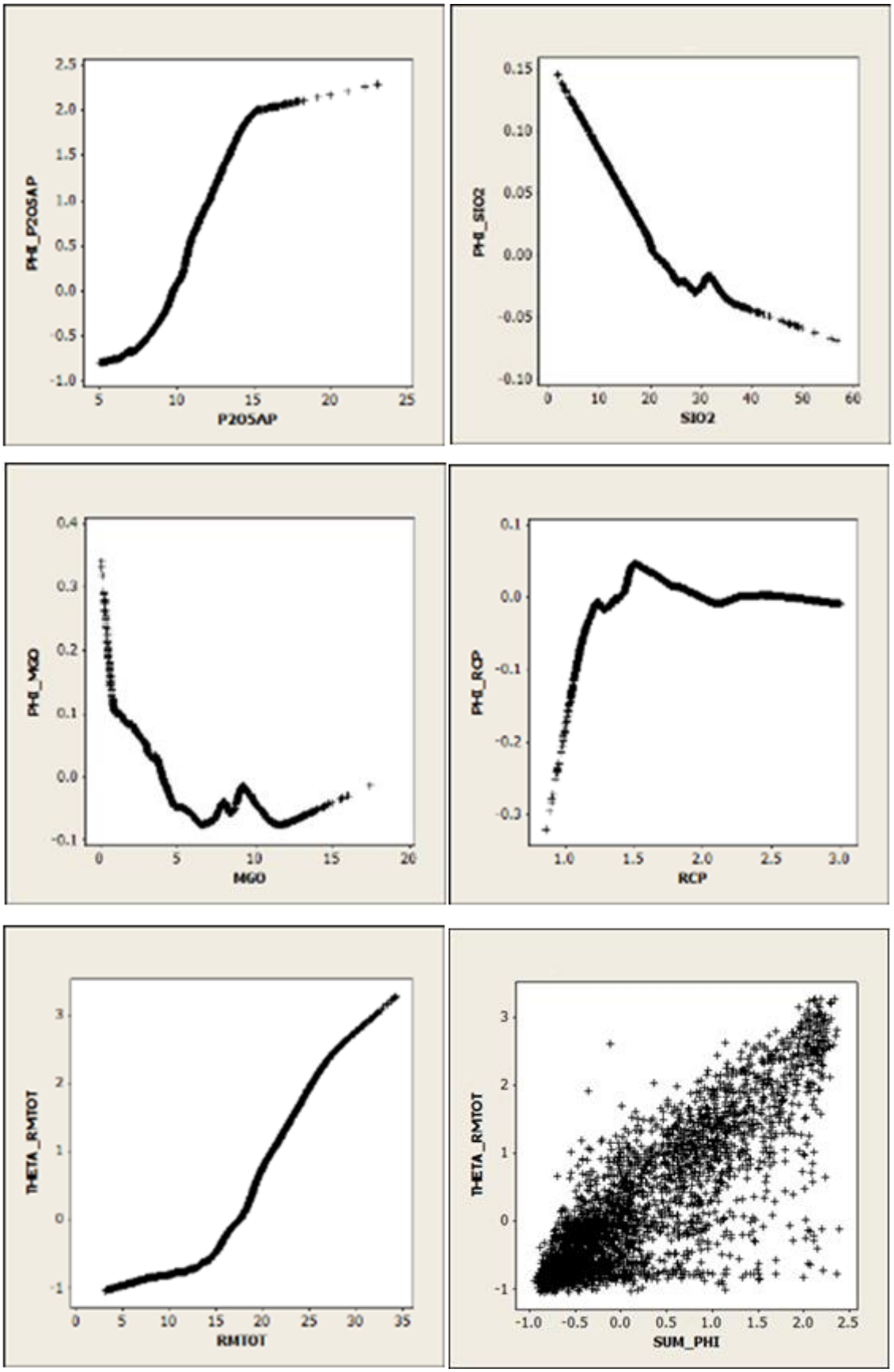


Figura 28 - Funções transformações *vs* variáveis - banco de dados faltante.

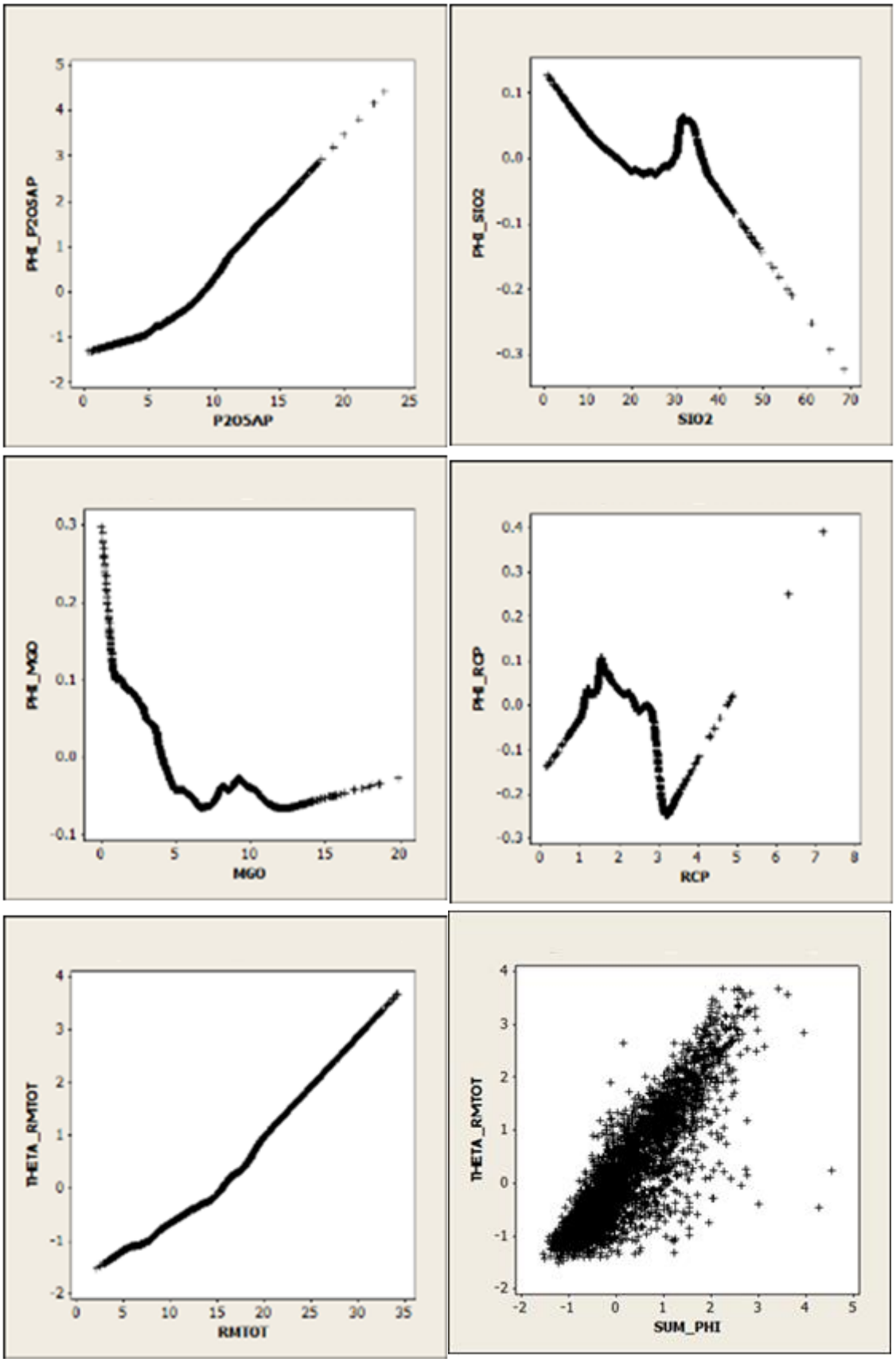


Figura 29 - Funções transformações *vs* variáveis - banco de dados cenário médio.

Pelos gráficos de dispersão de θ_RMTOT versus a soma das funções φ nota-se que no cenário médio os pontos estão mais próximos a reta de 45° e menos difusos que no cenário de dados faltantes, o que indica maior correlação e consequentemente um modelo mais acurado e preciso.

Após a definição interativa das funções de transformações para os 12 cenários em questão, as funções de saída do **ace.exe** (Barnett, 2013) são utilizadas no programa **ace_predict.exe** (Barnett, 2013) para prever a variável resposta (RMTOT) em um novo arquivo de variáveis preditoras obtendo os modelos geometalúrgicos finais. A figura 30 ilustra o arquivo de parâmetro do **ace_predict.exe** para o cenário médio, para os demais cenários é alterado o arquivo de entrada (ace_jmedia.out) e o arquivo de saída.

```
                Parameters for ACE_PREDICT
                **************************
START OF PARAMETERS:
ace_jmedia.out                        -file with original variables and ace transforms
-1.0e21     1.0e21                        - trimming limits
5 6                                       - cols for response and trans. response
4                                         - # of predictor variables
1 2 3 4                                  - cols for predictor variables
7 8 9 10                               - cols for predictor trans. predictors
data_amostras.txt                         -file with new predictor values
1 2 3 4                                  - cols for predictor variables
amostras_jmedia.out                           -output file for predicted response
```

Figura 30 - Arquivo de parâmetros do ace_predicit.exe

3.3.1 Regressão ACE – aplicação em pilhas de homogeneização

Para comparar os 12 modelos de regressão gerados, eles serão aplicados aos dados de formação de 23 pilhas de homogeneização, que alimentaram a usina de beneficiamento no período de 2016 e 2017.

A formação das pilhas de homogeneização da mina em estudo segue o seguinte fluxograma:

1. A partir do modelo de teores estimado por krigagem ordinária o programador de pilhas desenha as linhas (avanços de lavra) que limitam os blocos a serem lavrados;
2. Os avanços são lavrados e empilhados segundo modelo Chevron;
3. A cubagem desses avanços resulta nos teores e massas final da pilha, os quais são a base de informação para aplicar as regressões ACE, como pode ser visto na tabela 11.

Tabela 11 - Exemplo de programação de pilha de homogeneização

Avanços	Dados para Programação das Pilhas									
	ibilidade de	% P_2O_5 Apatítico	P_2O_5 Total	Fe_2O_3	MgO	SiO_2	TiO_2	Al_2O_3	CaO	RCP
	Máxima (t)	Calculado	(%)	(%)	(%)	(%)	(%)	(%)	(%)	
1	19 153	7.31	7.31	29.86	5.60	24.52	8.17	5.11	12.04	1.65
3	11 800	7.95	8.27	30.92	2.72	25.89	7.54	5.91	10.73	1.30
4	7 702	8.23	8.23	25.82	7.37	24.58	4.77	3.28	15.72	1.91
5	9 000	8.45	8.45	23.73	6.57	25.63	8.31	4.26	14.23	1.68
6	10 000	9.57	9.57	33.08	2.90	18.22	9.56	6.02	14.62	1.53
8	31 129	8.86	9.20	26.99	5.75	24.41	8.02	4.83	11.97	1.30
10	9 380	7.27	9.57	27.73	3.60	23.24	9.23	4.42	12.21	1.28
13	12 574	9.53	9.53	24.21	5.18	25.91	8.24	4.25	13.04	1.37
14	7 284	10.39	10.39	31.99	1.93	15.45	10.87	5.84	15.86	1.53
16	24 583	10.23	10.23	24.80	5.15	24.96	6.03	5.05	15.29	1.49
	142 605	8.97	9.17	27.67	5.10	23.08	7.94	4.90	13.76	1.50

Para garantir a não influencia de descontinuidade operacional no resultado de rendimento mássico de produção da usina nessas pilhas, foram selecionadas apenas dias que a usina teve um total de horas trabalhadas acima de 23h e 30 min e que não tenha havido troca de pilha alimentada à usina durante o período.

A Figura 31 apresenta os teores dessas pilhas versus suas respectivas recuperações mássicas. Nota-se, que as correlações entre a recuperação mássica obtida na usina de beneficiamento com os teores das pilhas, são menores do que as correlações observadas no banco de dados. Isso indica que mesmo selecionando períodos de estabilidade da usina, algumas variabilidades operacionais não podem ser medidas e mitigadas e serão consideradas como não influentes no resultado de recuperação mássica da planta. A tabela 12 apresenta as características médias das pilhas, onde fica evidente que as pilhas são formadas com teores similares, pois tem baixa variância em todas as variáveis de formação das pilhas. Com teores de alimentação médio de 8,56 % de P_2O_5ap, 23,54 % de SiO_2, 6,04 % de MgO e 1,7 % de RCP acarreta em uma recuperação mássica média de 14,27%.

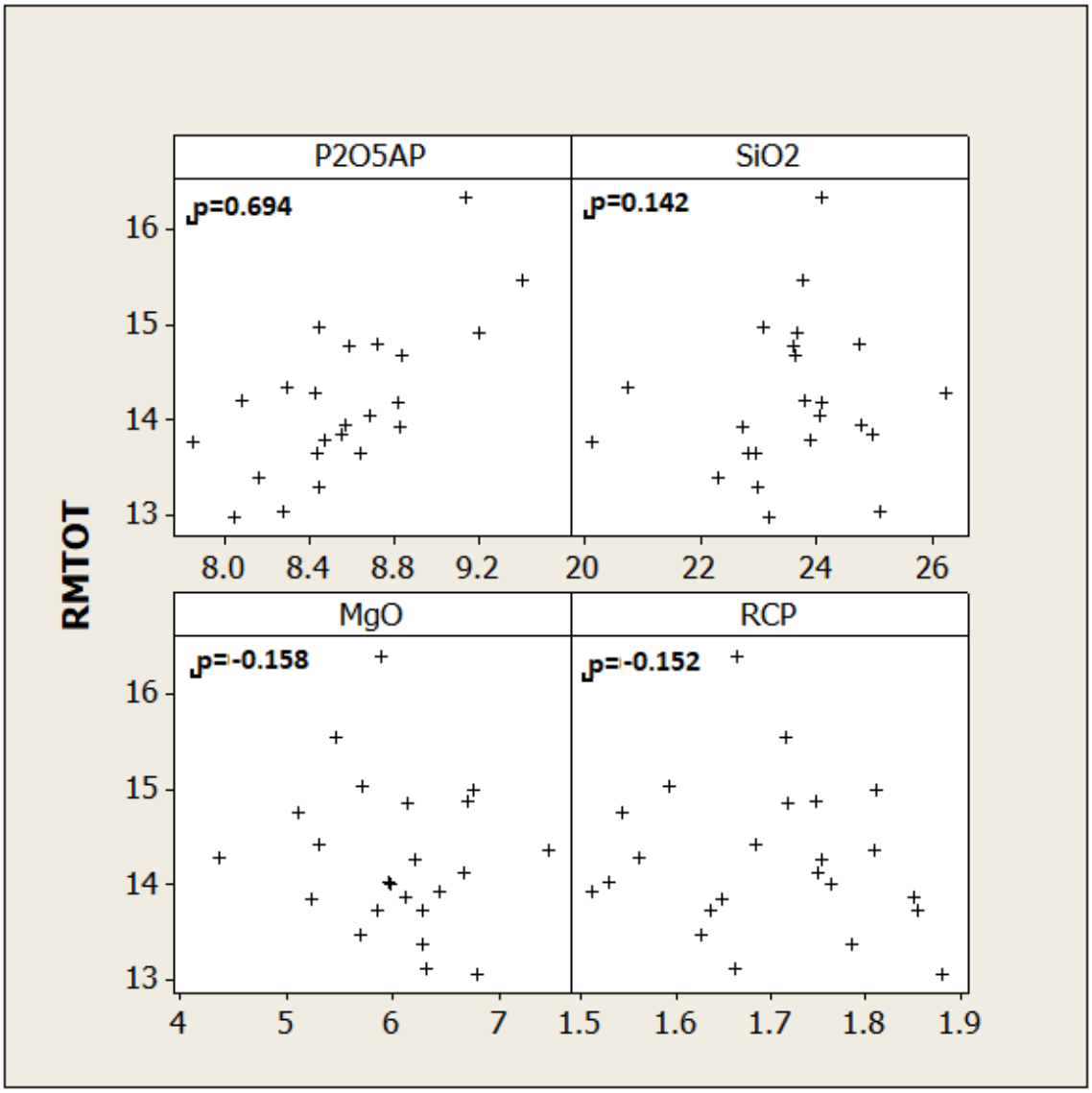


Figura 31 - Gráfico de dispersão dos teores *vs* RMTOT das pilhas que alimentaram à usina entre 2016 e 2017.

Tabela 12 - Estatística descritiva das pilhas de homogeneização processadas na usina de beneficiamento.

Variável	N⁰ de pilhas	Média	Mínimo	Máximo	Desvio pad.	Variância
P_2O_5ap	23	8,56	7,85	9,41	0,38	0,14
SiO_2	23	23,53	20,14	26,23	1,33	1,77
MgO	23	6,036	4,35	7,48	0,67	0,45
RCP	23	1,70	1,510	1,88	0,11	0,012
RMTOT usina	23	14,26	13,06	16,41	0,79	0,62

A figura 32 apresenta os gráficos de dispersão da aplicação dos modelos geometalúrgicos do cenário de dados faltantes e do cenário complementado com a média das múltiplas imputações versus o resultado na usina. Fica evidente, que o modelo com

dados faltantes superestimou os resultados da usina. O modelo com valores inseridos pelo valor médio obteve melhores resultados. Os gráficos dos 10 cenários encontram-se no anexo 2.

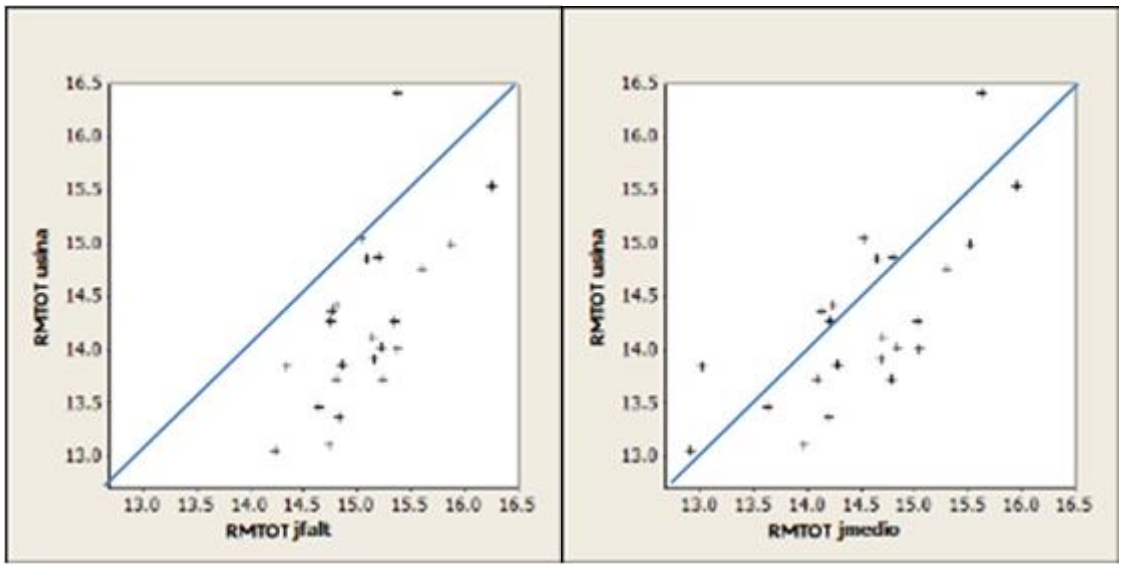


Figura 32 – Gráficos de dispersão do RMTOT usina *vs* RMTOT dos modelos jfalt e jmédio.

Pela análise de correlação dos 12 cenários (tabela 13), percebe-se que o cenário com os dados originais faltantes obteve a menor correlação com os resultados de usina, indicando que a falta de dados acarreta em menor precisão na previsibilidade da recuperação mássica do minério nas pilhas.

Tabela 13 - Coeficiente de correlação dos modelos geometalúrgicos dos 12 cenários *vs* resultado na usina das pilhas de homogeneização.

Correlação	RMTOT usina
RMTOT J1	0,757
RMTOT J2	0,749
RMTOT J3	0,723
RMTOT J4	0,756
RMTOT J5	0,74
RMTOT J6	0,702
RMTOT J7	0,734
RMTOT J8	0,726
RMTOT J9	0,761
RMTOT J10	0,702
RMTOT médio	0,732
RMTOT faltante	0,655

Para medir se houve viés na estimativa dos modelos geometalúrgicos, foi calculado o erro relativo de cada cenário, segundo:

$$Erro = \frac{RMTOTusina - RMTOTmodelo}{RMTOT\ usina} \quad (52)$$

A figura 33 apresenta os histogramas dos erros relativos dos 12 cenários, onde nota-se que a curva do cenário de dados faltantes (linha azul contínua) teve o erro relativo mais deslocado no sentido negativo do eixo x, o que indica uma superestimativa. O histograma do cenário médio (linha vermelha contínua) é o mais deslocado no sentido positivo do eixo x, com centro mais próximo de zero, indicando que foi o mais preciso.

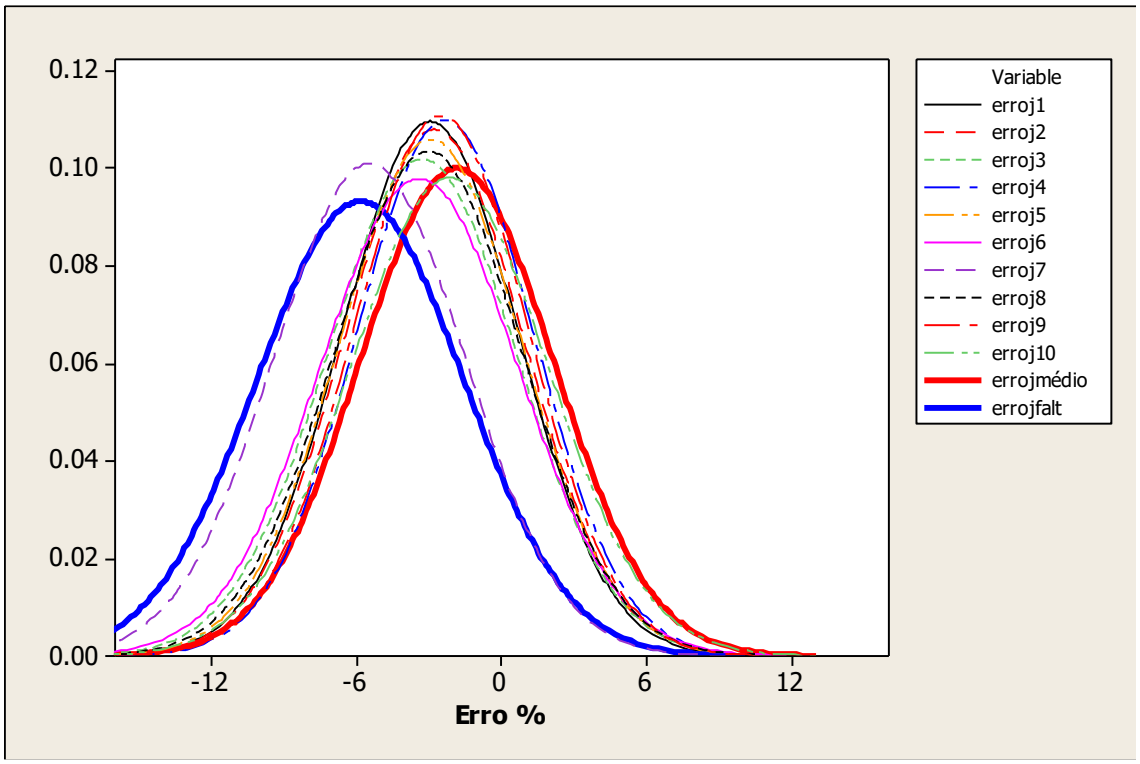


Figura 33 - Histograma dos erros relativos das pilhas nos 12 cenários.

A tabela 14 apresenta a estatística do erro, na qual se pode observar que além de ser o cenário com maior superestimativa, o cenário com dados faltantes também obteve maior variância indicando que foi a estimativa mais errática.

Tabela 14 - Estatística do erro relativo das pilhas nos 12 cenários.

Erro relativo (%)	N^0 de pilhas	Média	Mínimo	Máximo	Desvio pad.	Variância	IQR
erro1	23	-2,97	-8,062	5,001	3,643	13,269	5,095
erro2	23	-2,812	-7,803	5,162	3,698	13,672	5,201
erro3	23	-3,304	-8,836	4,499	3,918	15,353	6,393
erro4	23	-2,345	-7,406	5,576	3,637	13,224	4,679
erro5	23	-2,951	-8,105	5,056	3,771	14,222	5,113
erro6	23	-3,425	-9,701	4,54	4,08	16,649	5,748
erro7	23	-5,483	-10,917	4,053	3,954	15,635	6,202
erro8	23	-3,053	-8,685	5,15	3,859	14,893	5,122
erro9	23	-2,512	-7,506	5,351	3,606	13,002	5,349
erro10	23	-2,169	-7,932	5,431	4,074	16,597	6,715
erromedia	23	-1,87	-7,695	6,096	3,993	15,941	6,793
errofalt	23	-5,852	-12,356	6,344	4,283	18,34	6,106

A mina do estudo de caso utiliza em seu modelo de reservas para estimativa do RMTOT a metodologia de regressão linear definida por Fernandes (2013), apresentada na equação 53:

$$RMTOTreg.linear = 6.9 + (1.65 * P_2O_5) - (0.1 * Fe_2O_3) - (0.13 * TiO_2) - (0.13 * SiO_2) \quad (53)$$

No passado o modelo utilizado pela equipe de curto prazo considerava o RMTOT calculado a partir de um valor fixo de recuperação metalúrgica (RECTOT), segundo equação 54:

$$RMTOT\ calculado = (RECTOT * P_2O_5ap)/P_2O_5CON \quad (54)$$

Considerando o valor de RECTOT de 62,3 e de P_2O_5CON de 35,8.

Com objetivo de comparar essas duas metodologias que foram utilizadas nos modelos do depósito em questão à metodologia desenvolvida nessa dissertação, as metodologias de regressão linear (reg. Linear) e de cálculo a partir de valores fixos (calculada) foram aplicadas aos dados das 23 pilhas de homogeneização para obter os valores de RMTOT. Esses resultados foram comparados aos dados reais de RMTOT da usina. Pela tabela 15 que o método de regressão linear obteve erro relativo médio de 1,43 evidenciando uma pequena superestimativa. O método calculado obteve erro relativo médio de 4,56 o que representa uma maior superestimativa.

Tabela 15- Estatística do erro relativo das pilhas no modelo de regressão linear e calculado.

Erro relativo (%)	N^0 de pilhas	Média	Mínimo	Máximo	Desvio pad.	Variância	IQR
erro reg.linear	23	-1,433	-7,993	6,941	4,598	21,144	7,538
erro calculado	23	-4,561	-9,895	3,073	4,151	17,227	6,542

A figura 34 apresenta os gráficos de dispersão das duas metodologias com os respectivos valores de correlação obtidos entre os modelos e o resultado da usina. O

método de regressão linear obteve correlação de 0,616 e o método calculado correlação de 0,694.

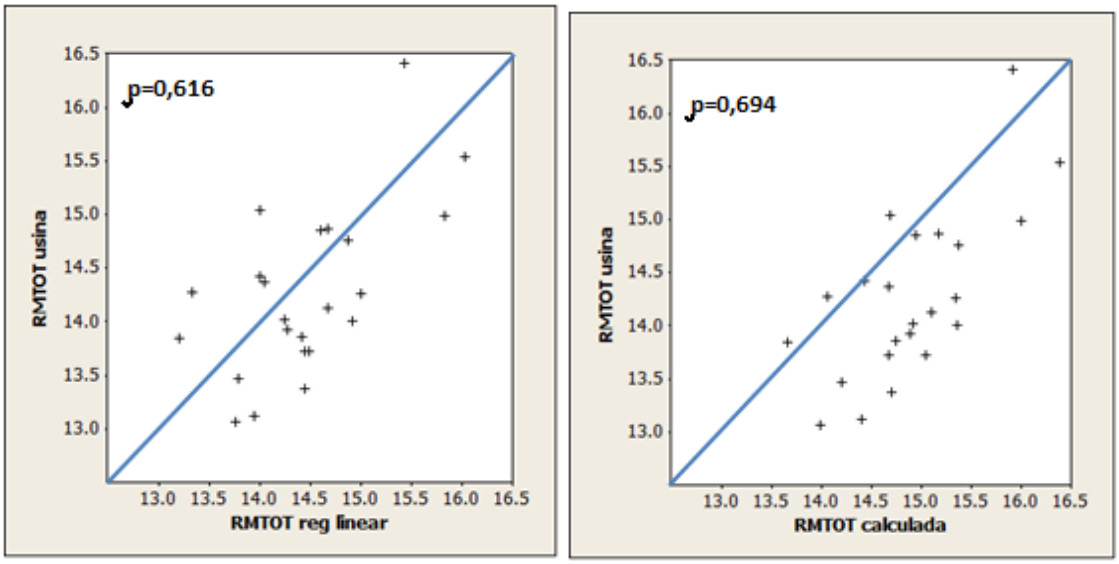


Figura 34 - Gráficos de dispersão do RMTOT usina *vs* RMTOT dos modelos regressão linear e calculado.

Comparando os dois métodos com o método desenvolvido no trabalho é possível concluir que em relação ao erro relativo, a regressão ACE com dados incompletos foi que obteve maior superestimativa, seguida pelo método de RMTOT calculado. O método de regressão ACE com dados completos obteve erro relativo muito próximo ao erro relativo obtido na regressão linear, que foi a metodologia com erro relativo mais baixo, indicando menor superestimativa. Porém ao comparar a correlação o método de regressão linear obteve o valor mais baixo de 0,616, seguido pela regressão ACE com dados incompletos com valor de 0,655. A metodologia que obteve maior correlação, ou seja, maior precisão na estimativa foi a regressão ACE com os dados completos.

Assim fica evidente que a regressão ACE com os dados isotópicos, completados por atualização Bayesiana, foi a metodologia que obteve melhor resultados.

3.3.2 Regressão ACE – aplicação em dados de laboratório

Para comparar os 12 modelos geometalúrgicos gerados por regressão ACE, desconsiderando qualquer influência de descontinuidades e ineficiências operacionais da usina de beneficiamento. Utilizaram-se os modelos aplicando-os aos novos resultados de testes em escala piloto. Foram selecionadas 131 amostras processadas nos meses de outubro e novembro de 2017, na planta piloto, para medir seu rendimento mássico.

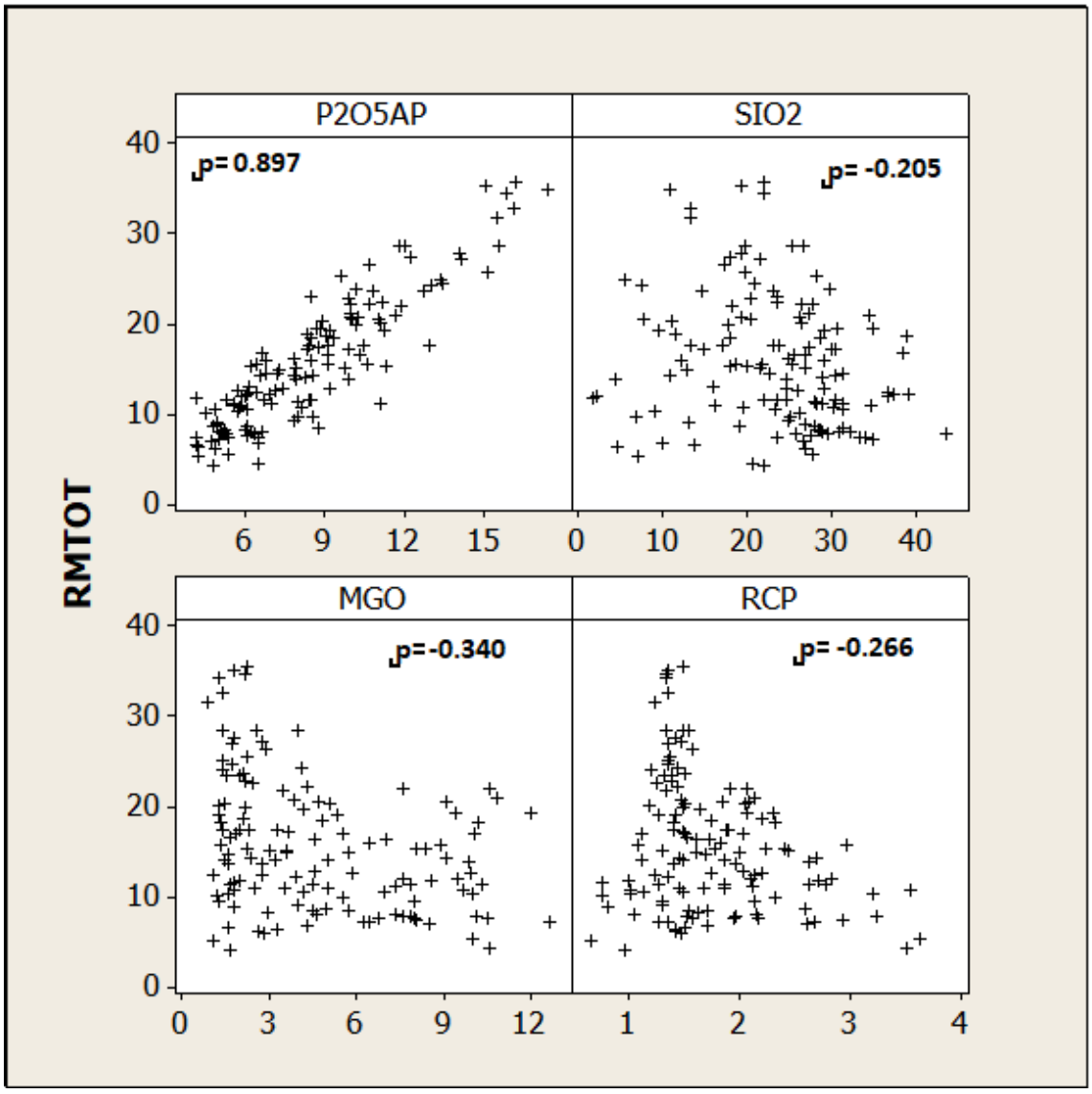


Figura 35 - Gráfico de dispersão dos teores *vs* RMTOT obtidos em testes de escala piloto.

Na figura 35, percebe-se que as correlações entre a recuperação mássica (RMTOT), obtida nas amostras em escala piloto, e seus teores são de 0,89 para P_2O_5ap, -0,205 para SiO_2, -0,34 para MgO e -0,266 para RCP. Essas correlações são similares as correlações observadas na tabela 1, obtidas entre a RMTOT e os teores do banco de dados de original. O que indica que essas novas amostras coletadas e analisadas na planta piloto obtiveram resultados aderentes aos das amostras do banco de dados do estudo. E isso ocorre, pois em escala laboratorial, não há instabilidades que possam influenciar no resultado de recuperação. Assim, ao aplicar os modelos de regressão geometalúrgicos para prever os resultados de RMTOT dessas novas amostras obtém-se valores que validam as metodologias utilizadas nessa dissertação, desconsiderando influências das oscilações causadas por ineficiência da usina.

A tabela 16 apresenta as estatísticas das novas amostras onde fica evidente, que diferentemente das pilhas, as amostras selecionadas têm uma variação maior nos teores, pois tem variâncias maiores. Indicando que a seleção das amostras foi aleatória, buscando considerar tanto os baixos como os altos teores. Os teores médios dessas amostras são de 8,57 % de P_2O_5ap, 23,13 % de SiO_2, 4,67 % de MgO e 1,75 % de RCP, bem similares aos teores de formação de pilha listados acima. A recuperação mássica média é de 15,74 %, ou seja, 1,5 % acima da recuperação média das pilhas.

Tabela 16 - Estatística descritiva das novas amostras.

Variavel	N⁰ de amostras	Média	Mínimo	Máximo	Desvio pad.	Variância
P_2O_5ap	131	8,569	4,19	17,41	3,05	9,33
SiO_2	131	23,13	1,72	43,59	8,52	72,65
MgO	131	4,67	0,89	12,69	3,14	9,89
RCP	131	1,74	0,66	3,63	0,59	0,34
RMTOT	131	15,73	4,13	35,59	7,25	52,60

A figura 36 apresenta os gráficos de dispersão da RMTOT obtidos pela aplicação dos modelos geometalúrgicos do cenário de dados faltante e do cenário complementado médio versus o resultado das RMTOT com as amostras novas em planta piloto. Percebe-se que não há viés significativo em nenhuma das estimativas, porém a distribuição dos dados no cenário médio é mais próxima à reta de 45° o que indica uma maior correlação. Os gráficos dos 10 cenários encontram-se no anexo 3.

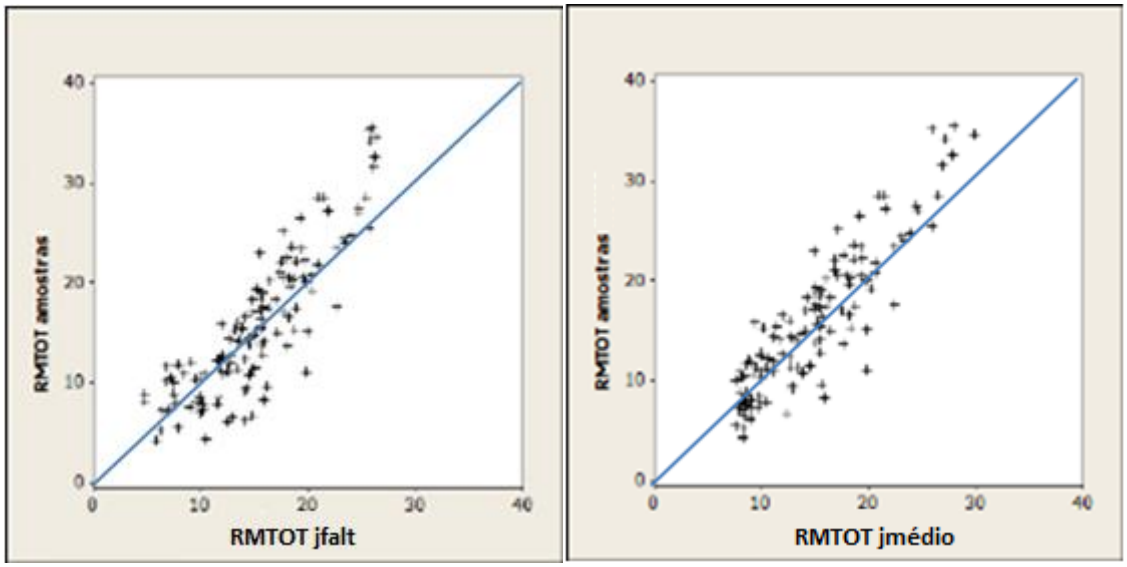


Figura 36 - Gráficos de dispersão das RMTOT obtidas com aplicação dos modelos jfalt e jmédio *vs* RMTOT com amostras novas na planta piloto.

Pela análise de correlação dos 12 cenários (tabela 17), pode-se comprovar que as previsões de RMTOT usando dados do cenário com os dados originais faltantes obteve a

menor correlação com os resultados de novas amostras em escala piloto. A falta de dados, mesmo para previsões em escala laboratorial de RMTOT, acarreta em menor precisão.

Tabela 17 - Coeficiente de correlação da RMTOT obtida por regressão com os 11 cenários complementado e com o cenário faltante X RMTOT com novas amostras em planta piloto.

Correlação	RMTOT amostras
RMTOT j1	0,9
RMTOT j2	0,899
RMTOT j3	0,9
RMTOT j4	0,899
RMTOT j5	0,899
RMTOT j6	0,9
RMTOT j7	0,898
RMTOT j8	0,9
RMTOT j9	0,9
RMTOT j10	0,896
RMTOT médio	0,901
RMTOT faltante	0,872

A figura 37 apresenta os histogramas acumulados dos erros relativos dos 12 cenários, onde nota-se que o cenário de dados faltantes (linha azul marinho) obteve o maior viés, superestimando os resultados de RMTOT.

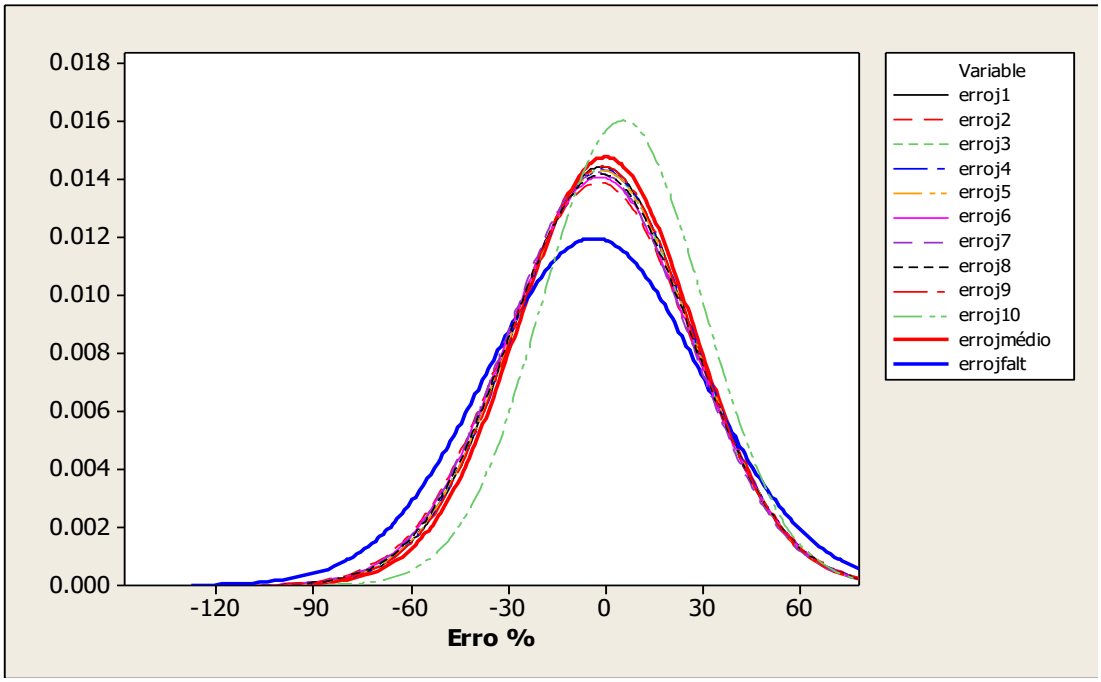


Figura 37 – Histograma do erro relativo das RMTOT com novas amostras em laboratório e RMTOT aplicado sobre os dados de laboratório aplicados aos modelos derivados dos 11 cenários de inserção e do cenário faltante.

A tabela 18 apresenta a estatística do erro, na qual pode-se comprovar que o cenário faltante obteve a maior superestimava e o cenário médio obteve o erro médio mais próximo de 0, ou seja não gerou viés nem positivo nem negativo.

Tabela 18 - Estatística do erro relativo das novas amostras laboratoriais nos 12 cenários.

Erro relativo (%)	Nº de amostras	Média	Mínimo	Máximo	Desvio pad.	Variância	IQR
erro1	131	-1,21	-101,36	38,2	27,58	760,87	29,14
erro2	131	-2,22	-112,73	37,61	28,7	823,56	30,71
erro3	131	-1,36	-107,28	36,56	27,82	773,84	27,62
erro4	131	-1,09	-106,5	39,24	27,8	772,76	27,82
erro5	131	-1,31	-108,41	39,3	27,84	774,86	33,33
erro6	131	-1,99	-118,83	36,61	28,33	802,55	29,34
erro7	131	-2,48	-106,64	35,2	28,02	785,03	36,04
erro8	131	-1,66	-109,93	36,95	28,12	790,58	28,21
erro9	131	-1,06	-101,18	37,63	27,57	760,16	28.00
erro10	131	5,05	-90,07	40,46	24,86	618,01	29,41
erromedio	131	-0,19	-100,9	40,88	26,99	728,24	29,1
errofalt	131	-3,79	-140,38	46,52	33,3	1108,73	30.00

Fica evidente, que assim como para as pilhas de homogeneização, para as amostras em escala de laboratório, os modelos de regressões dos cenários complementados obtiveram menores viés e maior precisão que o modelo de regressão com dados faltantes.

CAPÍTULO 4 CONCLUSÕES E RECOMENDAÇÕES

Um dos pontos cruciais na mineração é ter modelos precisos e acurados para previsibilidade do potencial de lucro. Para isso, é necessário o aproveitamento máximo dos recursos minerais, que só ocorre quando há conhecimento e controle geometalúrgico do depósito mineral. Desta forma, é possível prever a produção em escala industrial do minério.

O modelo geometalúrgico, que engloba variáveis geológicas e metalúrgicas é a base para conhecer o potencial do bem mineral na usina de beneficiamento. Na maioria das operações de mina, a análise em planta piloto da variável resposta metalúrgicos é inexistente, ou quando ocorre é significativamente menor que o número de amostras químicas. Geralmente, a disposição espacial da informação é concentrada apenas nas regiões de alto teor, o que pode acarretar em viés e inacuracidade ao modelo.

O estudo em questão sugeriu que essa incompletude dos dados metalúrgicos, principalmente quando a falta de dados é não aleatória (MNAR), no caso vinculada aos baixos teores, deve ser tratada previamente para poder gerar um modelo que seja representativo do depósito.

No estudo de caso realizado em uma mina de fosfato, os resultados faltantes de recuperação mássica (RMTOT), vinculados aos baixos teores foram completados por atualização bayesiana com transformação fixa, a fim de se obter um banco de dados completo respeitando o mecanismo de falta MNAR. Com esse banco de dados completo, foi gerado um modelo geometalúrgico por regressão ACE, a qual tem como característica a minimização do erro e maximização da correlação entre as variáveis.

Para verificar se houve ganho de precisão, quando os dados originais com amostras faltantes foram complementados, foram gerados modelos geometalúrgicos, por regressão ACE, com o cenário original faltante e com os cenários complementados. Esses modelos foram gerados utilizando os resultados de RMTOT e teores de 23 pilhas de homogeneização processadas na usina e com os resultados de RMTOT vindo de 131 novas amostras processadas em planta piloto.

4.1 CONCLUSÕES – ESTUDO DE CASO

A reconciliação das pilhas processadas na usina demonstrou que a inserção de dados por atualização bayesiana, tornando o banco de dados completo, melhorou a acuracidade e precisão das estimativas do modelo geometalúrgico. O cenário com os

valores médios de múltiplas inserções usados para prever RMTOT obteve correlação de 0,732 com os resultados medidos; enquanto, que o cenário com dados originais incompletos obteve correlação de 0,65. Além disso, o erro relativo do cenário médio foi de -1,65 contra -5,85 no cenário com dados originais. Indica que a falta de teores baixos causou uma superestimativa dos valores de RMTOT com o modelo geometalúrgico derivado, comprovando que ignorar a falta de amostras em um banco de dados leva a estimativas enviesadas.

Quando comparadas as metodologias de modelagem geometalúrgica do RMTOT utilizados na mina em estudo: regressão linear de Fernandes (2013), RMTOT calculado a partir da recuperação metalúrgica fixa e a regressão ACE gerada a partir dos dados completos, apresentada nessa dissertação. Esta última foi a metodologia que obteve maior precisão e acuracidade para previsão do RMTOT das pilhas de homogeneização.

A aplicação da regressão ACE para prever RMTOT nos dados de ensaios em planta piloto (novas amostras) obteve correlações da ordem 0,8 para todos os cenários, incluindo o cenário com banco de dados incompleto. Porém, esse último também obteve a mais baixa correlação e a maior superestimativa. Devido as altas correlações encontradas e erros relativos próximos de 0 para essas regressões ACE com amostras laboratoriais, pode-se concluir que a regressão mostrou uma alta aderência para geração do modelo geometalúrgico.

4.2 RECOMENDAÇÕES

Como análise final, ficam algumas considerações e recomendações:

A reconciliação do modelo geometalúrgico com as pilhas de homogeneização na usina de beneficiamento apresentou uma correlação da ordem de 73%, que é considerada baixa. Indica que, é necessário avaliar se não há outros fatores influenciando no potencial de recuperação do minério em escala industrial.

Na confecção de modelos geometalúrgicos, sugere-se sempre avaliar a influência de dados faltantes e quando necessário trabalhar para sua completude, de forma a não gerar viés em qualquer regressão que se deseja construir a partir de um conjunto de dados onde parte dos mesmos são faltantes.

REFERÊNCIAS

ALLISON, P. D., *Missing data*. Thousand Oaks: A Sage University Paper, 2002. 93p.

BALTAR, C. A. M., *Flotação no tratamento de minérios*. Recife: Editora Universitária da UFPE, 2010, 238 p.

BARNETT, R. M., *Tools for multivariate geostatistical modeling*. Edmonton: University of Alberta, Centre for computational Geostatistics, 2011, 97p.

BARNETT, R. M. e DEUTSCH, C. V., Multivariate imputation of unequally sampled geological variables. Houston: Mathematical Geoscience, v. 47, n. 7, pp.791-817, 2015.

BARNETT, R. M. e DEUTSCH, C. V., Tutorial and tools for ace regression and transformation. Edmonton: CCG Annual report 15, 2013, Paper 401.

BLISS, C. L., The method of probits. Washington: Science, v. 79, n. 2037, pp. 38-39, 1934.

BREIMAN, L. e FRIEDMAN, J. H., Estimating optimal transformations for multiple regression and correlation. Alexandria, VA: Journal of the American Statistical Association, v. 80, n. 391, pp. 580-598, 1985.

BROD, J. A., RIBEIRO, C. C., GASPAR, J. C., JUNQUEIRA-BROD, T. C., BARBOSA, E. S. R., RIFFEL, B. F., SILVA, J. F., CHABAN, N., FERRARI, A. J. D., Excursão 1. geologia e mineralizações dos complexos alcalino- carbonatíticos da província ígnea do alto Paranaíba. In: Congresso brasileiro de geologia, 42., 2004, Araxá, MG. Anais Recursos minerais e desenvolvimento socioeconômico, Belo Horizonte: SBG, 2004. pp. 1-29.

BROWNLEE, K. A., Statistical theory and methodology in science and engineering. New York: John Wiley & Sons, 1967. 590p.

BULATOVIC, S. M., Handbook of flotation reagents: chemistry, theory and practice flotation of sulfide ores. Ontario, Canada: Elsevier Science, 2007, 458p.

COHEN, J., Statistical power analysis for behavioural sciences. New York: Erlbaum publishers, 1988. 559p.

DEMIRTAS, H. e SCHAFER, J. L., On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. Statistics in Medicine, v.22, n.16, pp. 2553–2575, 2003.

DEMPSTER, A. P., LAIRD, N. M. e RUBIN, D. B., Maximum likelihood from incomplete data vie the EM algorithm, New York: Journal of the Royal Statistical Society, v. 39, n. 1, pp. 1-38, 1977.

DEUTSCH, C. V. e JOURNEL, A. G., Gslib: a geoestatistical software library and user's guide. NewYork: Oxford University Press, 1998. 363p.

DOYEN, P. M., DEN BOER, L. D. e PILLET, W. R., Seismic porosity mapping in the ekofisk field using a new form of collocated cokriging, 1996. In: SILVA, C. Z., Metodologias de inserção de dados sob mecanismo de falta MNAR para modelagem de teores em depósitos multivariados heterotópico. 2018. 96f. Tese (Doutorado em Engenharia) – Programa de Pós-Graduação em Engenharia de Minas, Metalurgia, e de Materiais, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2018.

ENDERS, C. K., Applied missing data analysis. New York: The Guilford Press, 2010. 410p.

FÁVERO, L. P., BELFIORE, P., SILVA, F. L. e BETTY, L. C., Análise de dados – modelagem multivariada para tomada de decisões. Rio de janeiro: Campus/Elsevier, 2009. 672p.

FERNANDES, F. G., Estudo do melhor método de extrapolação de regressão múltipla para construção do modelo geometalúrgico de uma mina de fosfato brasileira. 2013. 130f. Dissertação (Mestrado em Engenharia de Minas) – Programa de Pós – Graduação em Engenharia Mineral, Universidade Federal de Ouro Preto, Ouro Preto, 2013.

GOOVAERTS, P., Geostatistics for natural resource evalution. New York: Oxford University Press, 1997. 483p.

GREEN, P. J. e SILVERMAN, B. W., Nonparametric regression and generalized linear models: a roughness penalty approach. Londres: Chapman and Hall, 1994. 184p.

GUIMARÃES, R. C., ARAÚJO, A. C., PERES, A. E. C., Reagents in igneous phosphate ores flotation. Minerals Engineering, v.18, pp. 199-204, 2005.

HÄRDLE, W., Applied nonparametric regression. Londres: Cambridge University Press, 1992. 352p.

HARTMANN, R. S., A monte carlo analysis of alternative estimators in models involving selectivity. Londres: Journal of Business and Economic Statistics, v. 9, n. 1, pp. 41-49, 1991.

HECKMANN, J. T., Geostatistics for natural resource evaluation. New York: Oxford University Press, 1979. 483p.

ISAAKS, E. H. e SRIVASTAVA, R. M., An introduction to applied geoestatistics, New York: Oxford University Press, 1989. 561p.

JOHNSON, R. A., WICHERN, D. W. Applied Multivariate Statistical Analysis. Pearson, 2002. 808p.

JOURNEL, A. G., Nonparametric estimation of spatial distributions. Journal of the International Association for Mathematical Geology, v. 15, n. 3, pp. 445-468, 1983.

KRUSKAL, J. B., Analysis of factorial experiments by estimating monotomic transformations of the data. New York: Journal of Royal Statistical Society, B 27, pp. 251-263, 1965.

LANDIM, P. M. B., Análise estatística de dados geológicos, Rio Claro – SP: Editora Unesp, 2003. 256p.

LEAL, R. S., PERONI, R. L., COSTA, J. F. C. L., PEREIRA, S.G., MARTINS, R. R., CAPPONI, L. N., Geostatistics applied to geometallurgical modeling. In: 24th World Mining Conference, 24., 2016, Rio de janeiro. Anais Mining in a World of Innovation. pp.126-132.

LEAL FILHO, L. S., CHAVES, A. P., Flotação, 2004. In: LUZ, A. B., SAMPAIO, J. A., MONTE, M. B. M., ALMEIDA, S. L. M., Tratamento de Minérios. Rio de Janeiro: CETEM/MCT, 2004. 858p.

LEJA, J., Surface chemistry of froth flotation. New York: Plenum Press, 1982. pp. 611-665.

MONTE, M. B. M. e PERES, A. E. C., Química de superfície na flotação, 2004. In: LUZ, A. B., SAMPAIO, J. A., MONTE, M. B. M., ALMEIDA, S. L. M., Tratamento de Minérios. Rio de Janeiro: CETEM/MCT, 2004. 858p.

NGUYEN, A. V., SCHULZE, H. J., Colloidal science of flotation. New York: CRC press, 2004. 840p.

NGUYEN CONG, V. e RODE, B. M., Application of alternating condicional expectations method to quatitative electronic structure-activity relationships (QESAR). Molecular Informatics Journal, v.14, n.16, pp. 512-517, 1995.

PERES, A. E. C. e ARAÚJO, C., Flotação como operação unitária no tratamento de minérios, 2006. In: CHAVES, A. P., Teoria e prática do tratamento de minérios, Volume 4, A flotação no Brasil. São Paulo: Signus Editora, 2009, 29p.

PERSECHINI, M. A. M., JOTA, F. G., OLIVEIRA, M. L. M., PERES, A. E. C., Instrumentação de uma coluna de flotação piloto para desenvolvimento de técnica de controle avançada. Rio de Janeiro: CETEM – Série Tecnologia Mineral, 2001. 40p.

PUHANI, P. A., The heckman correlation for sample selection and its critique. Journal of Economic Surveys, v.14, n. 1, pp. 53-67, 2000.

RAO, S. R., Surface chemistry of froth flotation. New York: Kluwer Academic/Plenum Publishers, 2004, pp. 385-478.

REN, W., Exact downscaling in reservoir modeling. 2007. 213f. Tese de Doutorado. Centre of Computational Geostatistics, University of Alberta, Edmonton, 2007.

RODRIGUEZ, P. C., GRACIOSO, J. E., CABALEIRO, S. M. O. L., Utilização de conceitos geometalúrgicos para elaboração de cartas de controle. In: IBRAM, seminário nacional, 4, 1990. Anais: O computador e sua aplicação no setor mineral: pesquisa, lavra e beneficiamento minerais".

ROSSI, M. e DEUTSCH, C. V., Mineral resource estimation. New York: Springer, 2014. 332p.

RUBIN, D. B., Multiple imputation for nonresponse in surveys. New York: John Wiley and Sons Ltda, 1987. 253p.

RUBIN, D. B., Multiple imputations in sample surveys – a phenomenological bayesian approach to nonresponse. Journal of the American Statistical Association, pp. 20-34, 1978.

RUBIN, D. B., Inference and missing data. Biometrila, v. 63, n. 3, pp. 51-592, 1976.

SARMA, D. D., Geoestatistics with applications in earth science. India: Springer, 2009. 206p.

SCHAFER, J. L. E GRAHAM, J. W., Missing data: our view of the state of the art. Psychological Methods, v.7, n. 2, pp. 147-177, 2002.

SCHOFIELD, C.G., Homogenisation/Blending systems design and control for mineral processing. TransTech Publucations, Germany, 1980. 236p.

SILVA, C. Z., Metodologias de inserção de dados sob mecanismo de falta MNAR para modelagem de teores em depósitos multivariados heterotópico. 2018. 96p. Tese (Doutorado em Engenharia) – Programa de Pós-Graduação em Engenharia de Minas, Metalurgia, e de Materiais, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2018.

SINCLAIR, A. J. E BLACKWELL, G. H., Applied mineral inventory estimation. Londres: Cambridge University Press, 2004. 400p.

VERLY, G. W., 1984, Estimation of spatial point and block distributions: the multigaussian model. 1984. 416f. Tese de Doutorado, Stanford University, Stanford, 1984.

VIEIRA, M. C. A., 2016, Metodologia para prever recuperação de zinco em planta de beneficiamento. 2016. 154f. Dissertação (Mestrado em Engenharia) - Programa de Pós-Graduação em Engenharia de Minas, Metalurgia, e de Materiais, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016.

WALLER, L. A. E GOTWAY, C. A., Applied spatial statistics for public health data. Nova Jersey: Wiley – Interscience, 2004. 502p.

WACKERNAGEL, H., Multivariate geoestatistics: an introduction with applications. Berlim: Springer, 1998. 338p.

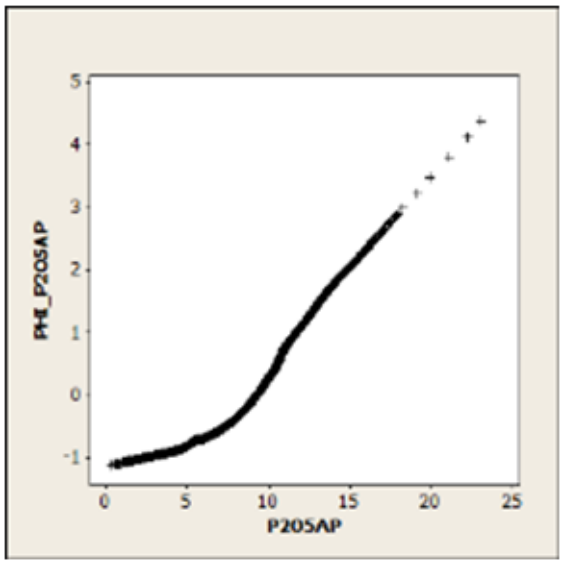
WANG, D. E MURPHY, M., Estimating optimal transformations for multiple regression using ace algorithm. Journal of Data Science, v. 2, pp. 329-246, 2004.

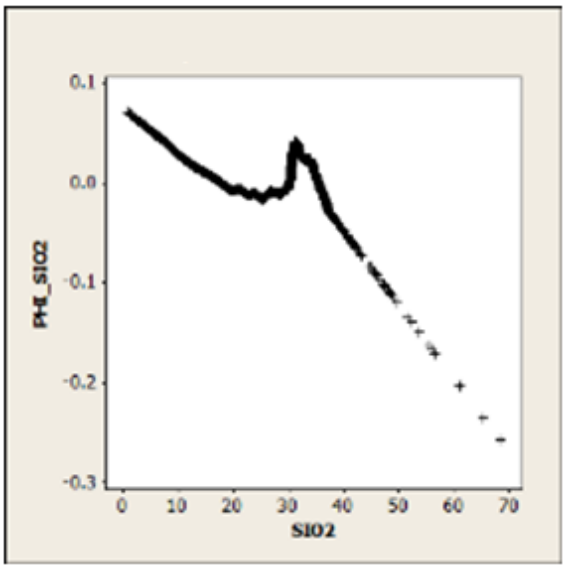
WILLIAMS, S. R. e RICHARDSON, J. M., Geometallurgical mapping: a new approach that reduces technical risk. In: 36th Annual Meeting of the Canadian Mineral Processors, 36, 2004, Paper 16.

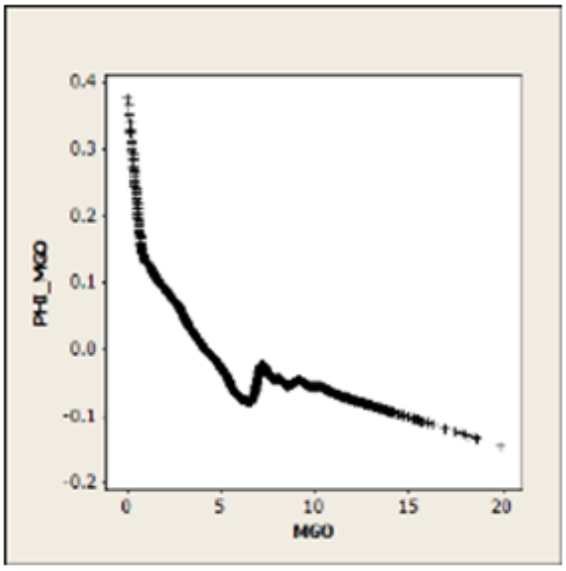
WILLS, B. A. e NAPIER-MUNN, T. J., Mineral processing technology: an introduction to the practical aspects of ore treatment and mineral recovery. Oxford: Butterworth-Heinemann, 2007. 465p.

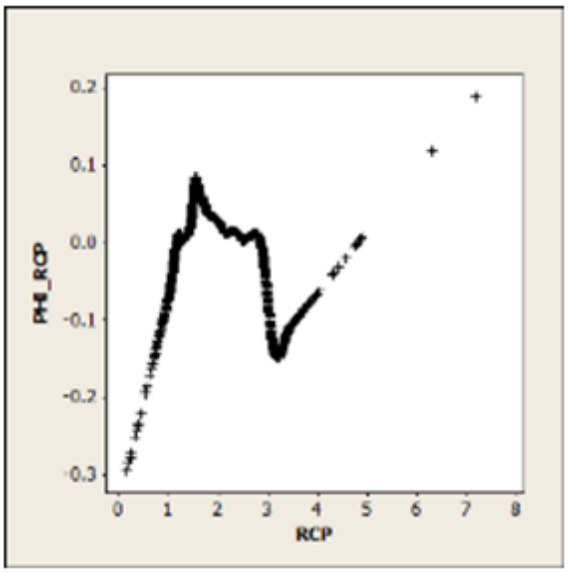
ANEXO 1: FUNÇÕES TRANSFORMAÇÕES *VS* VARIÁVEIS

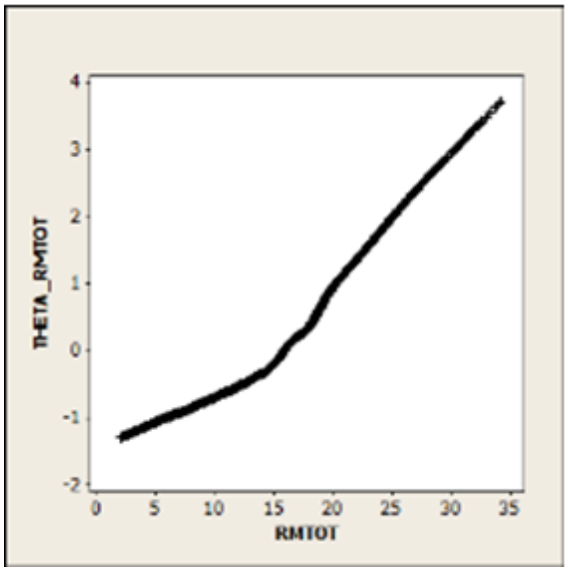
- Cenário J1:

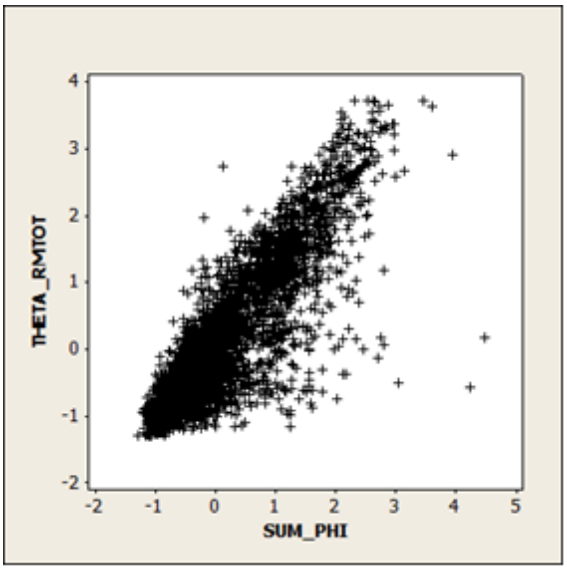




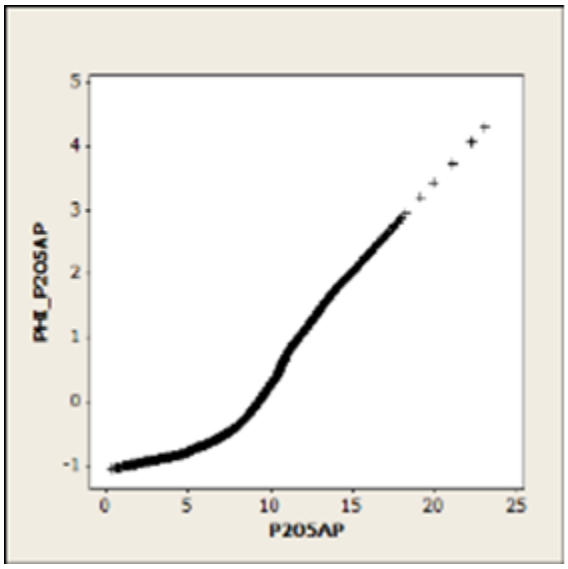


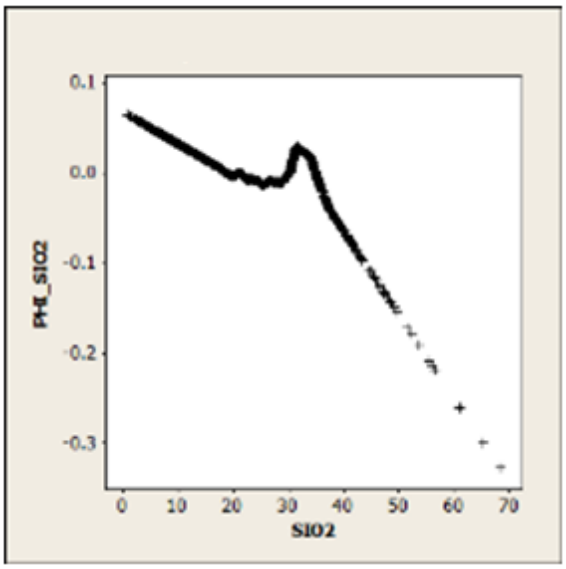


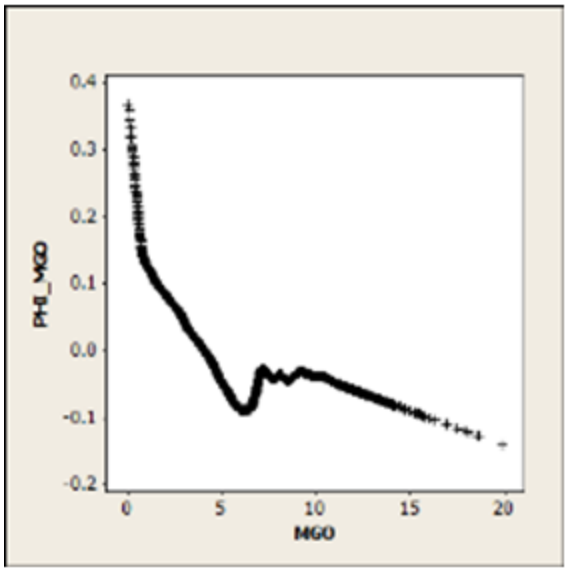


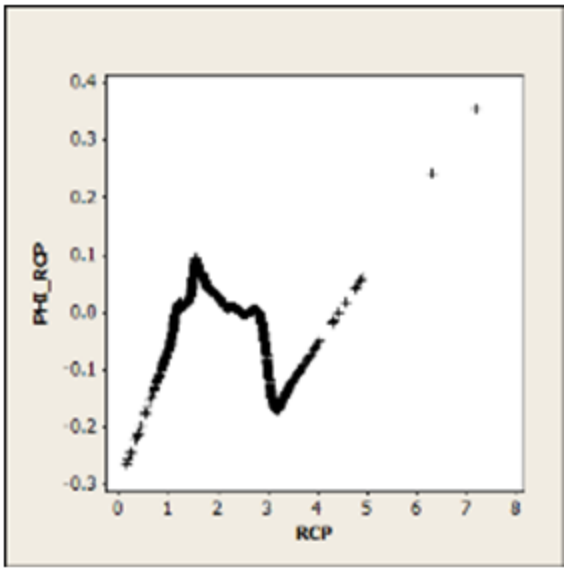


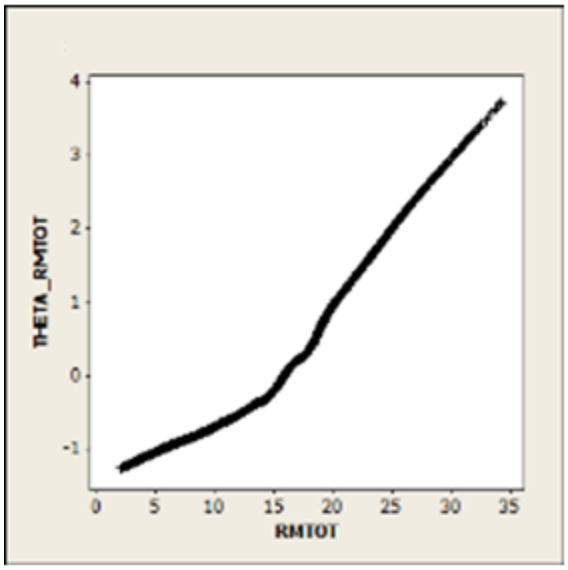
- Cenário J2:

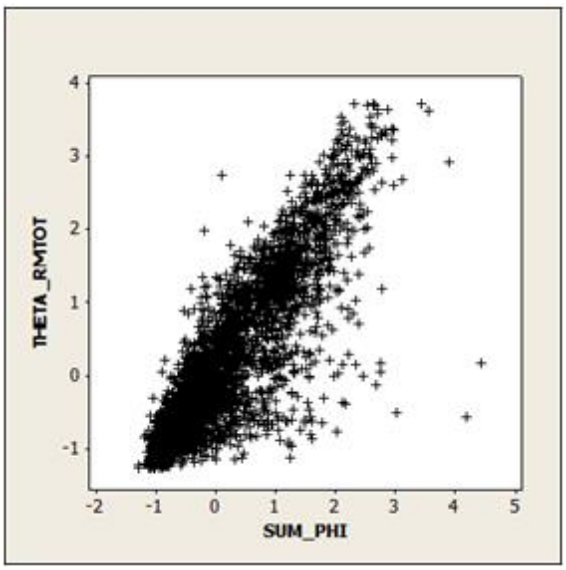

PHI_P2O5AP
P2O5AP


PHI_SIO2
SIO2

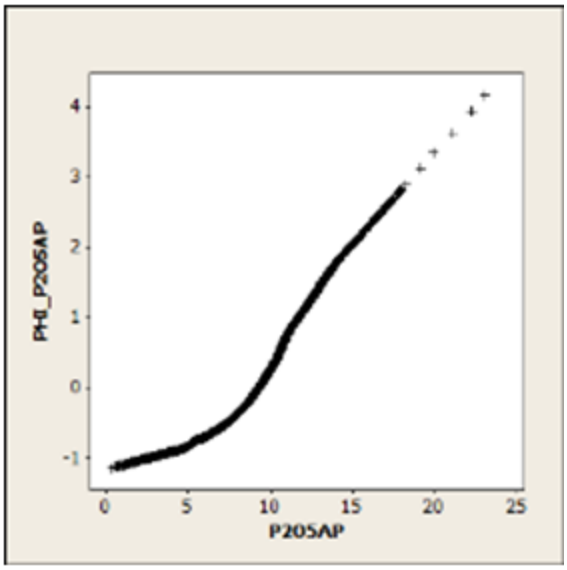

PHI_MGO
MGO

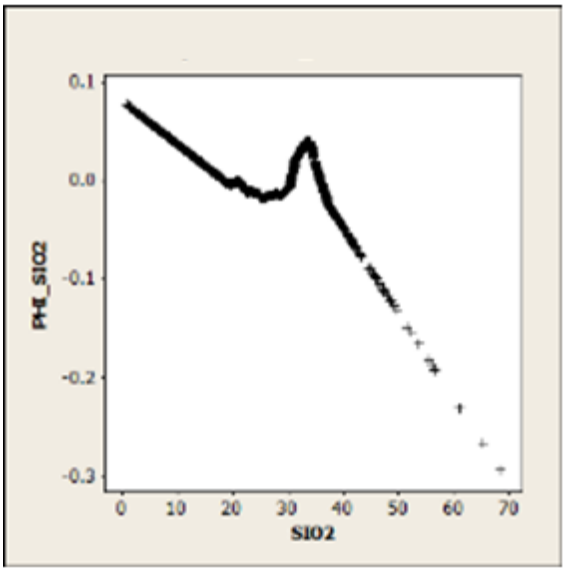

PHI_RCP
RCP

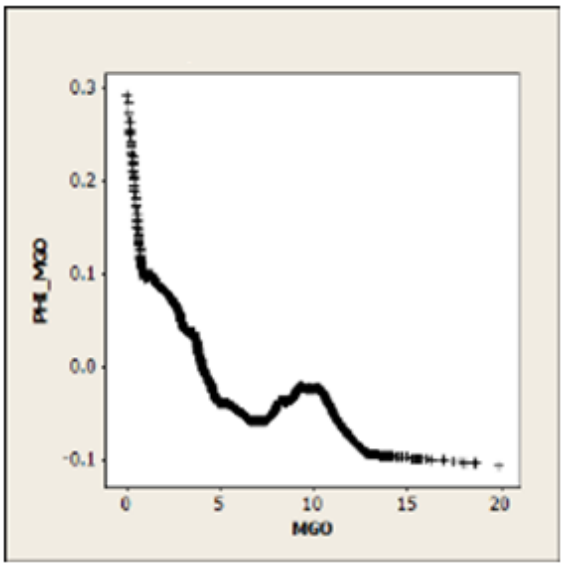

THETA_RMTOT
RMTOT

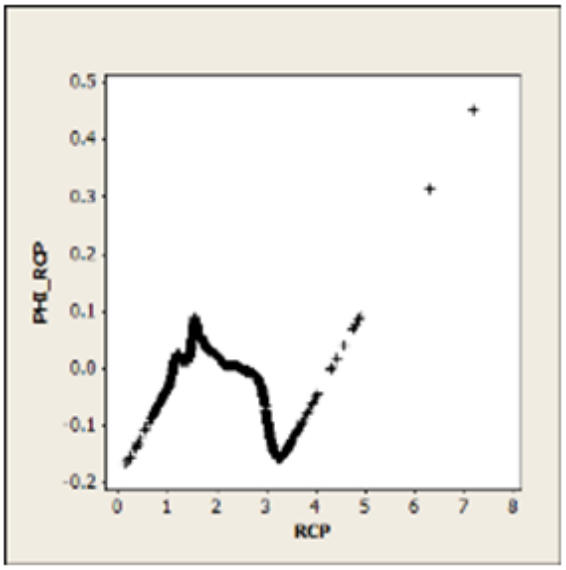

THETA_RMTOT
SUM_PHI

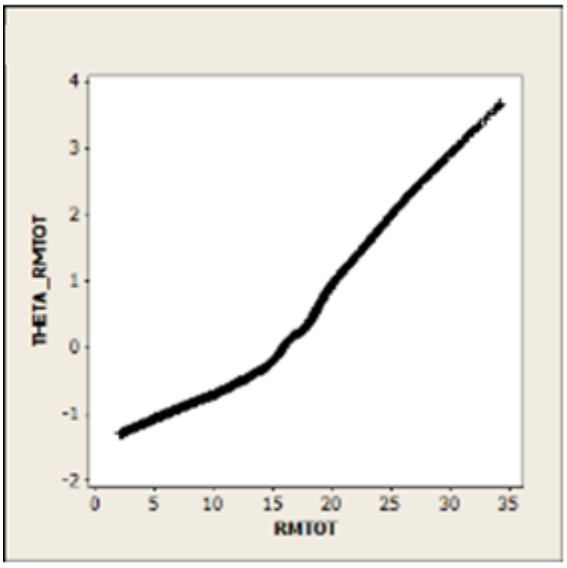
- Cenário J3:

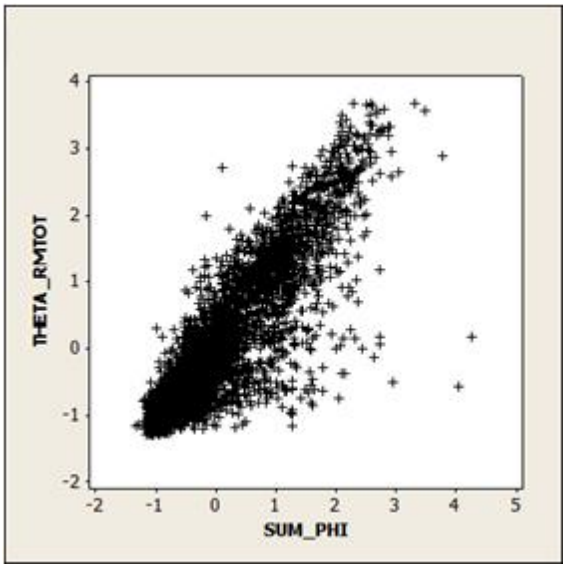

PHI_P2O5AP
P2O5AP


PHI_SIO2
SIO2

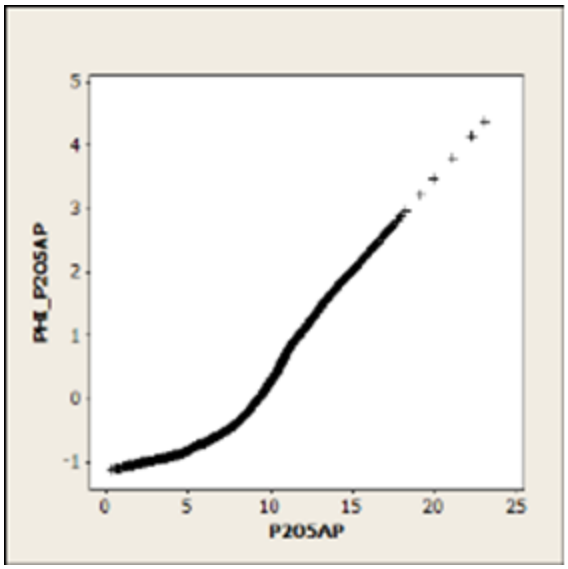

PHI_MGO
MGO

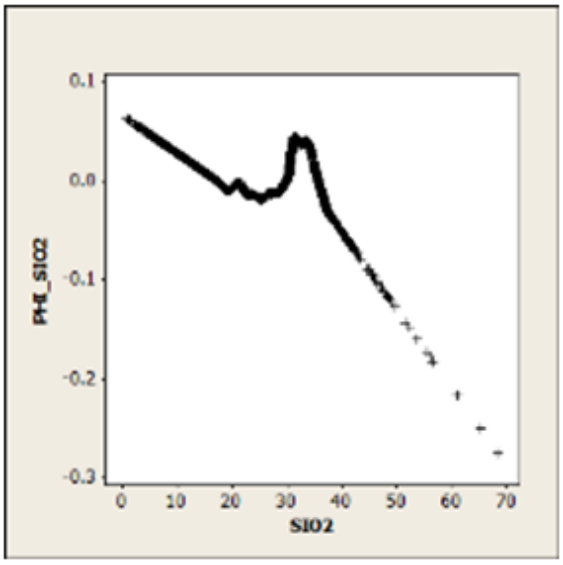

PHI_RCP
RCP

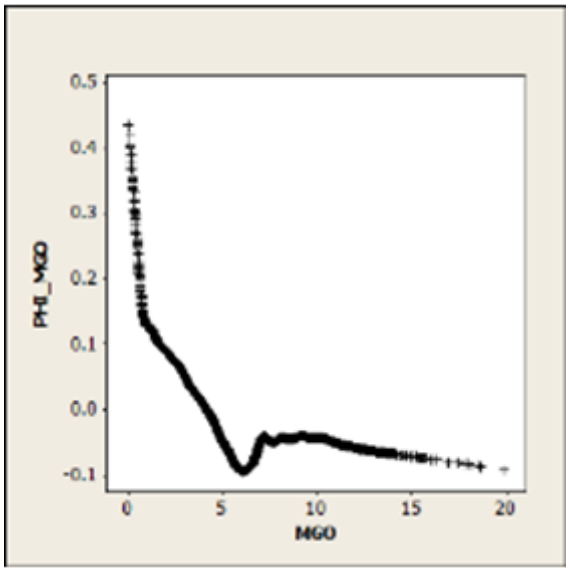

THETA_RMTOT
RMTOT

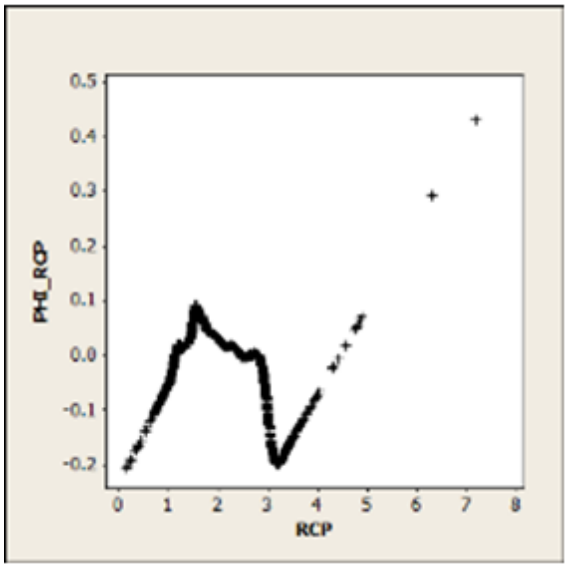

THETA_RMTOT
SUM_PHI

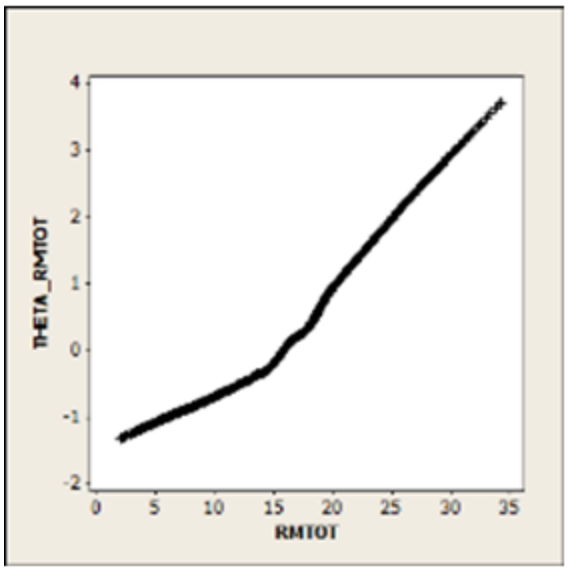
- Cenário J4:

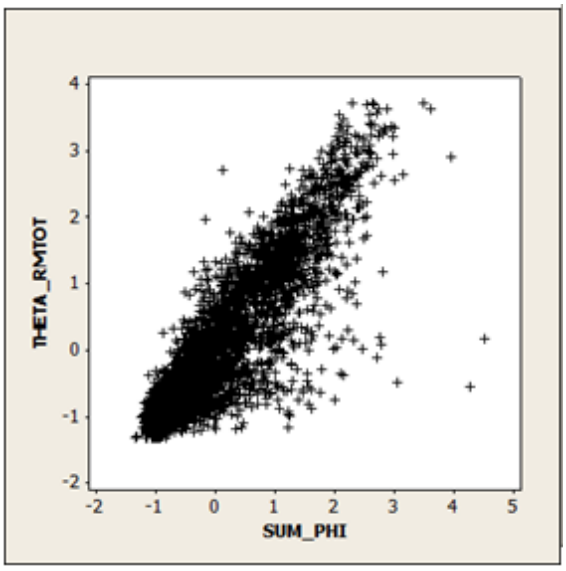

PHI_P2O5AP
P2O5AP


PHI_SIO2
SIO2

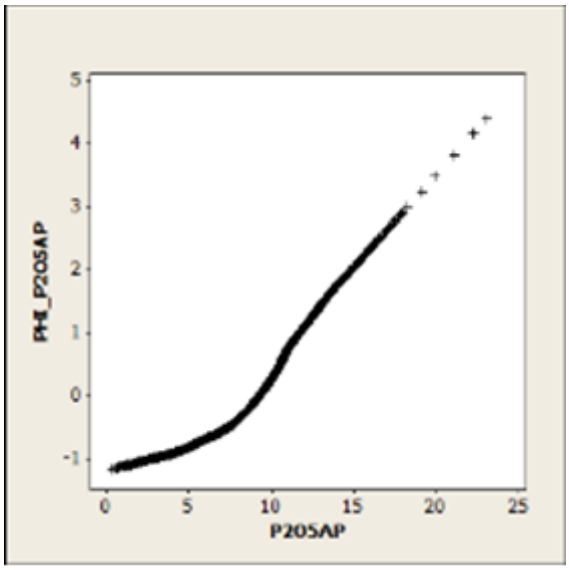

PHI_MGO
MGO

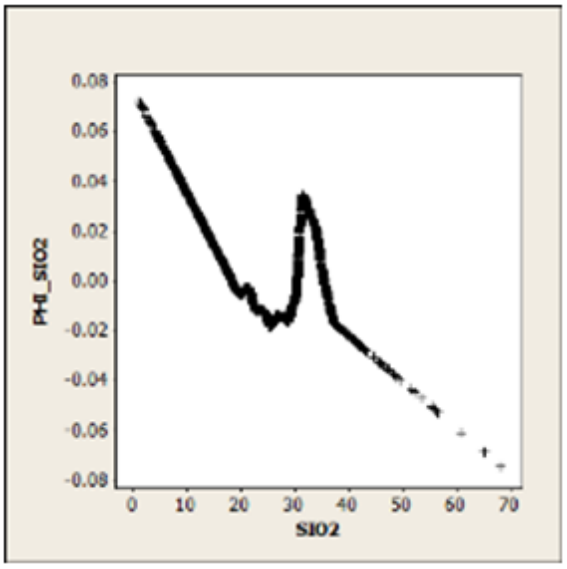

PHI_RCP
RCP

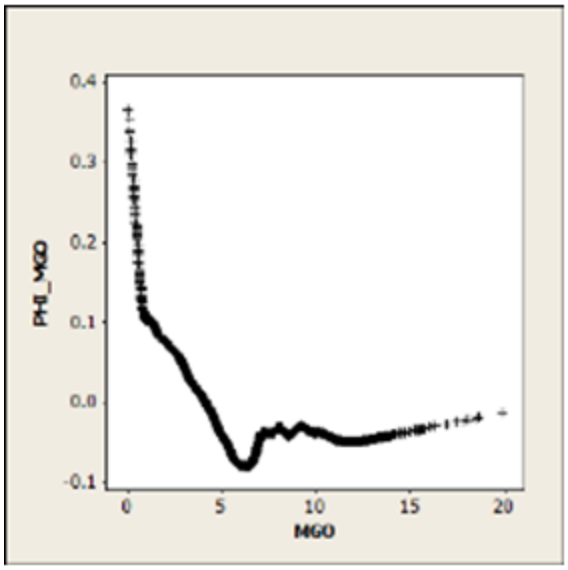

THETA_RMTOT
RMTOT

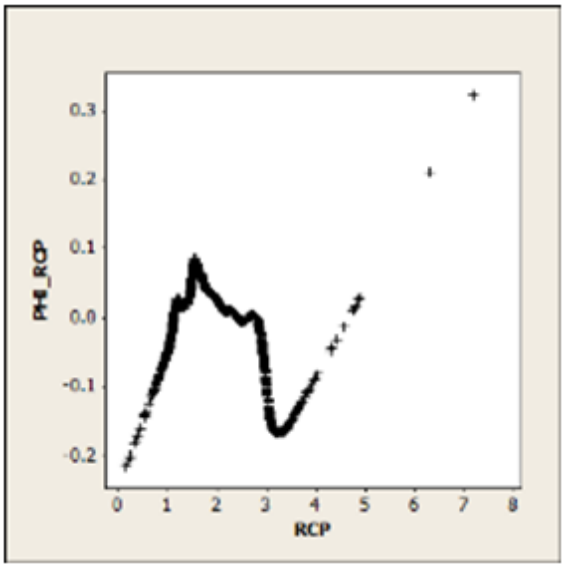

THETA_RMTOT
SUM_PHI

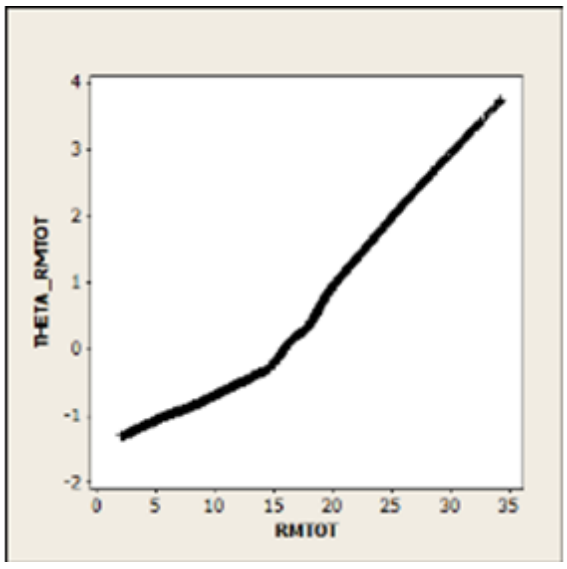
- Cenário J5:

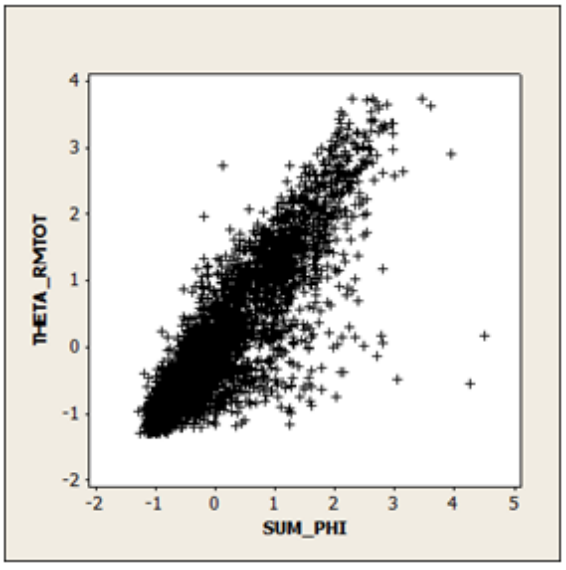

PHI_P2O5AP
P2O5AP


PHI_SIO2
SIO2

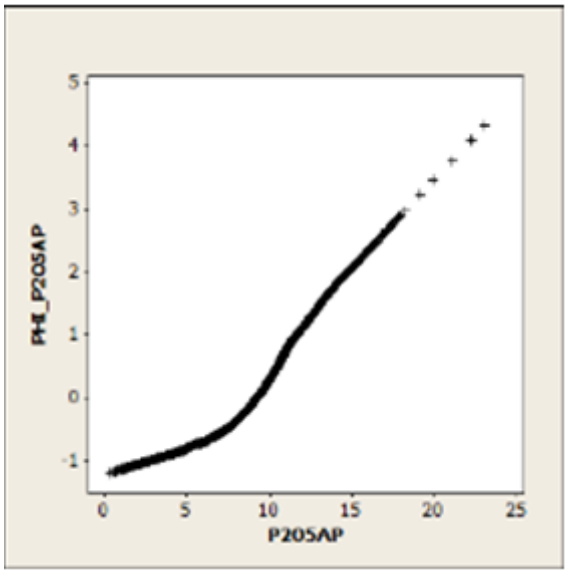

PHI_MGO
MGO

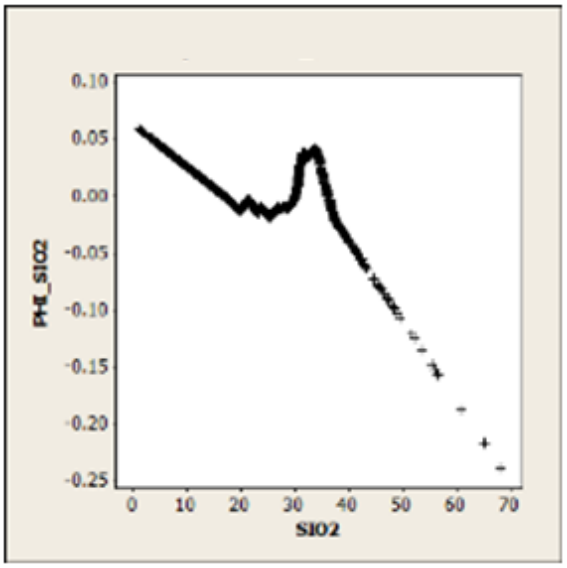

PHI_RCP
RCP

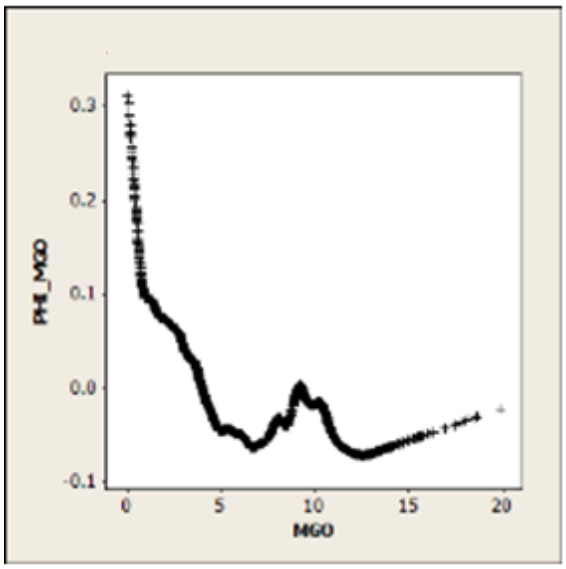

THETA_RMTOT
RMTOT

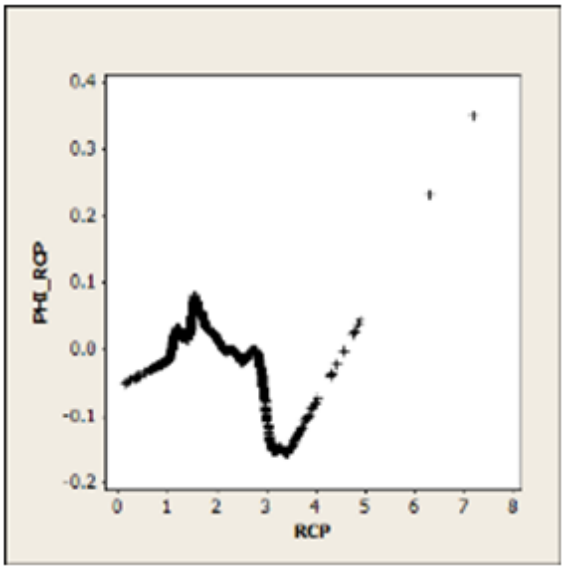

THETA_RMTOT
SUM_PHI

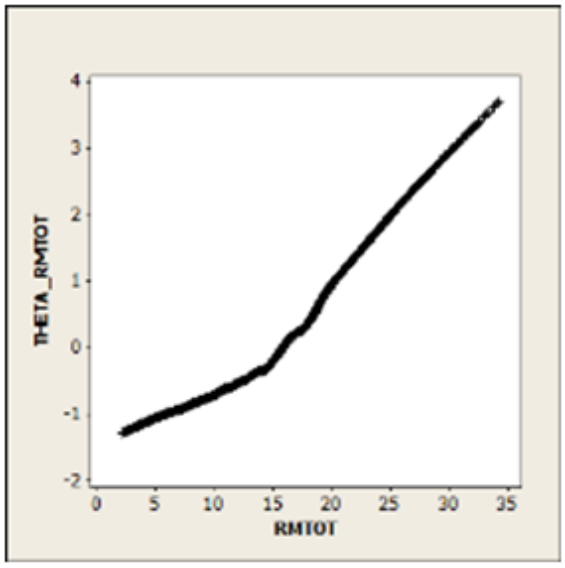
- Cenário J6:

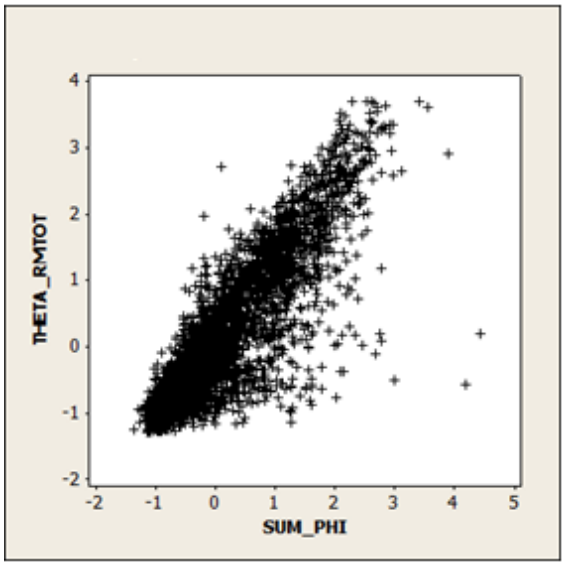

PHI_P2O5AP
P2O5AP


PHI_SIO2
SIO2

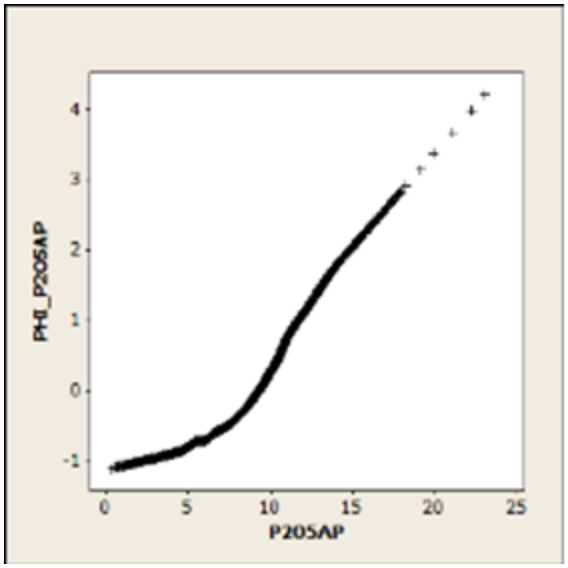

PHI_MGO
MGO

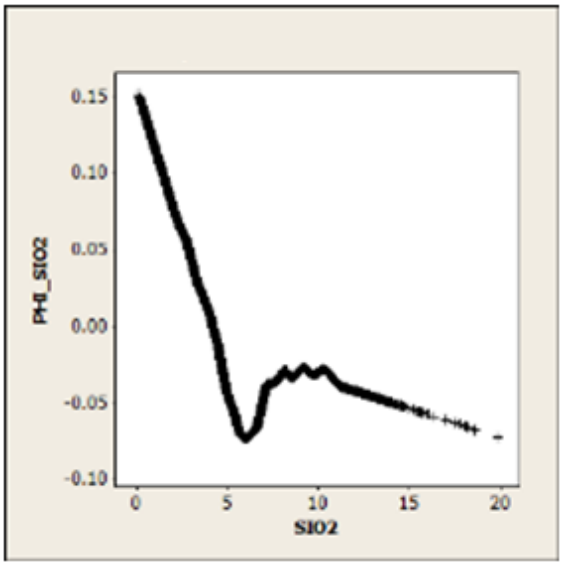

PHI_RCP
RCP

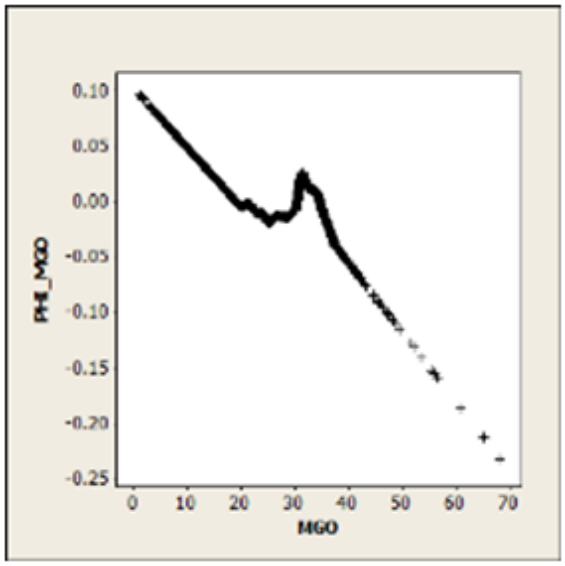

THETA_RMTOT
RMTOT

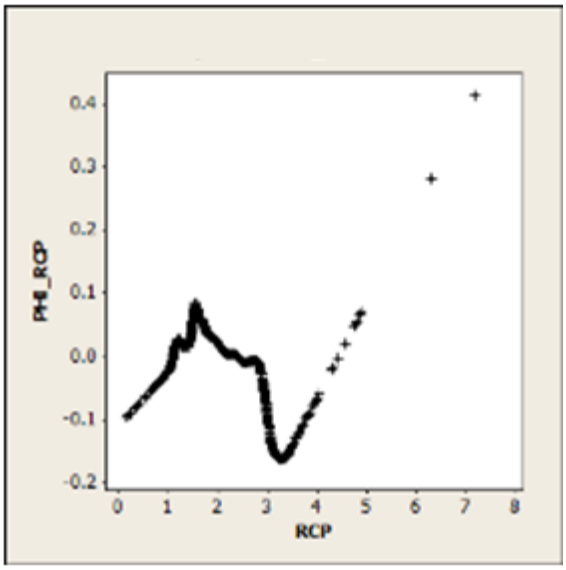

THETA_RMTOT
SUM_PHI

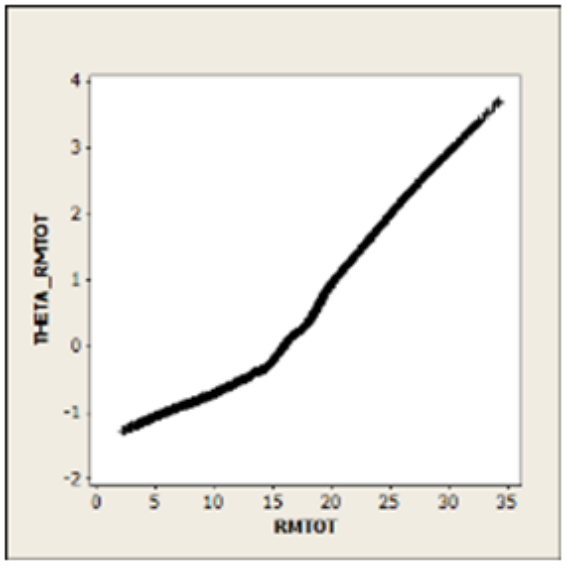
- Cenário J7:

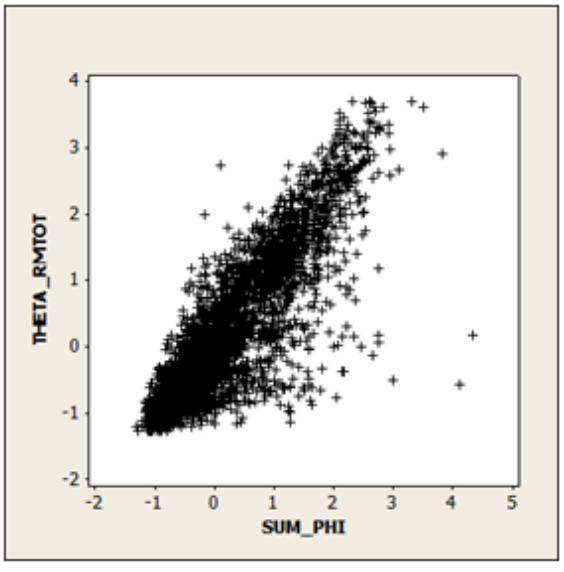

PHI_P2O5AP
P2O5AP


PHI_SIO2
SIO2

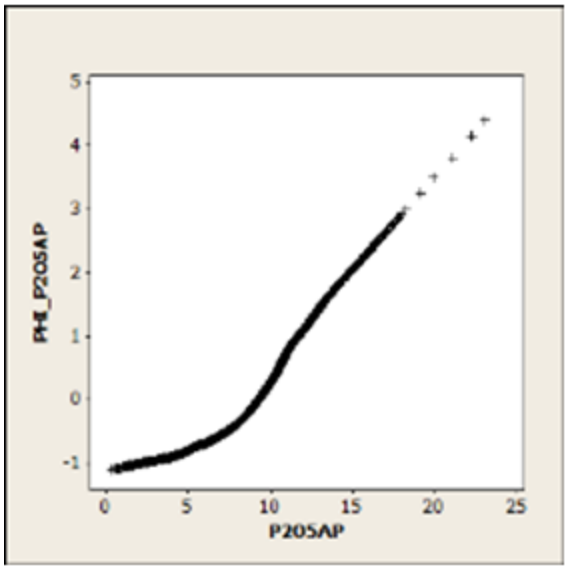

PHI_MGO
MGO

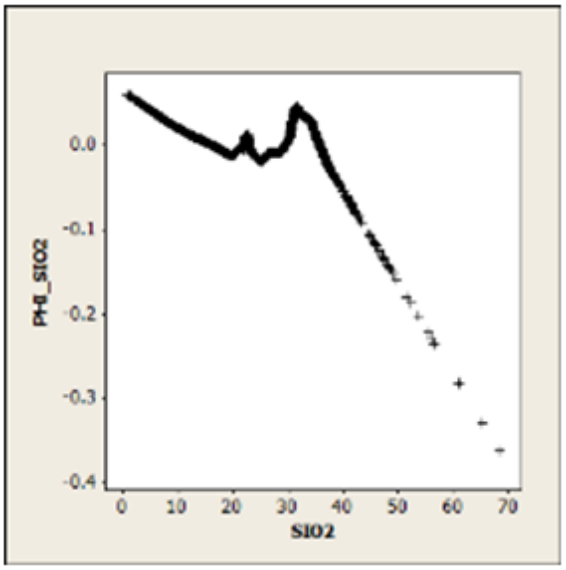

PHI_RCP
RCP

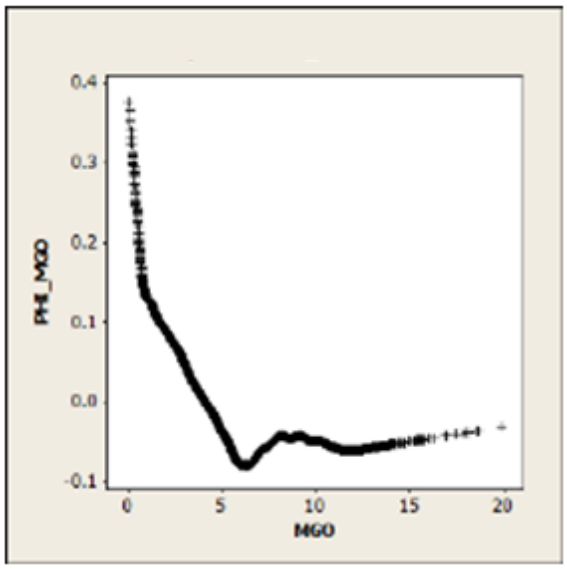

THETA_RMTOT
RMTOT

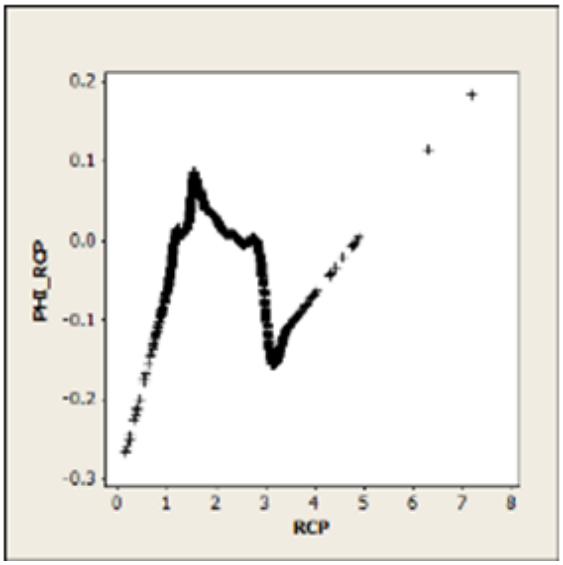

THETA_RMTOT
SUM_PHI

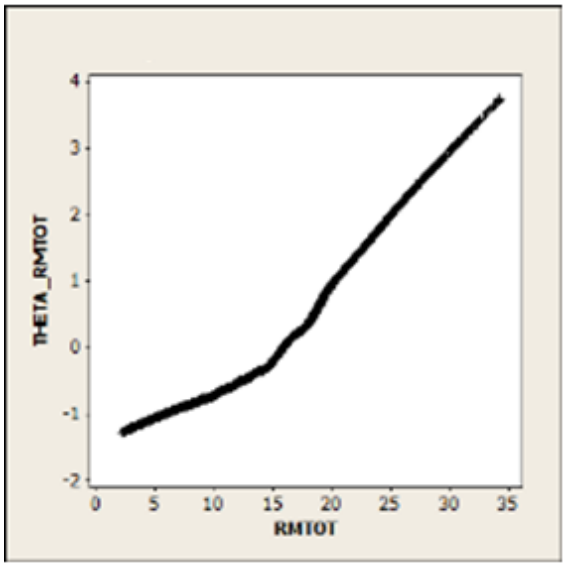
- Cenário J8:

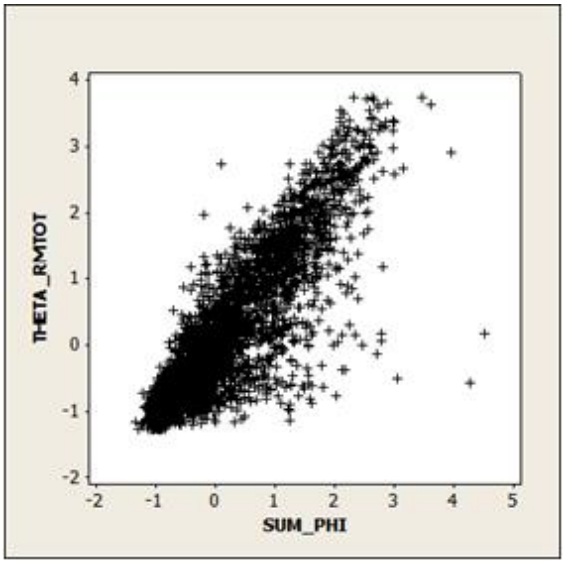

PHI_P2O5AP
P2O5AP


PHI_SIO2
SIO2

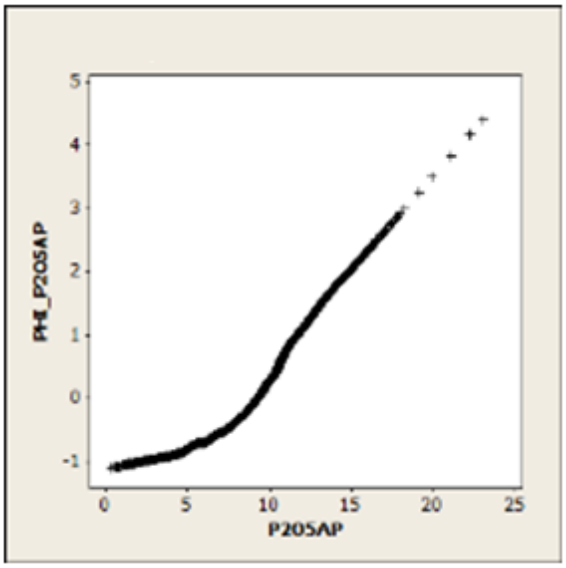

PHI_MGO
MGO

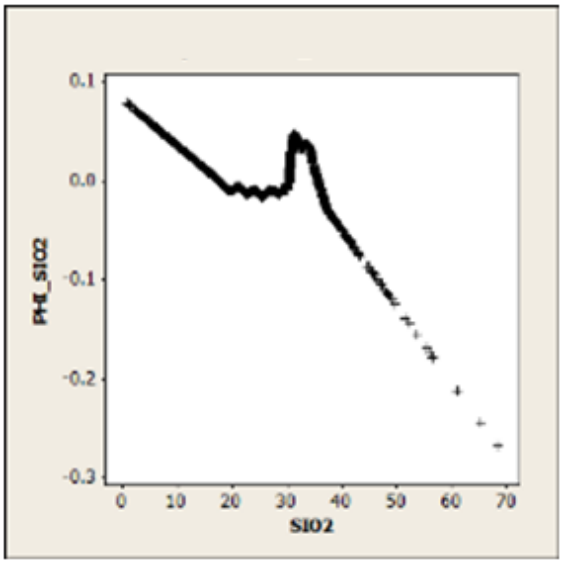

PHI_RCP
RCP

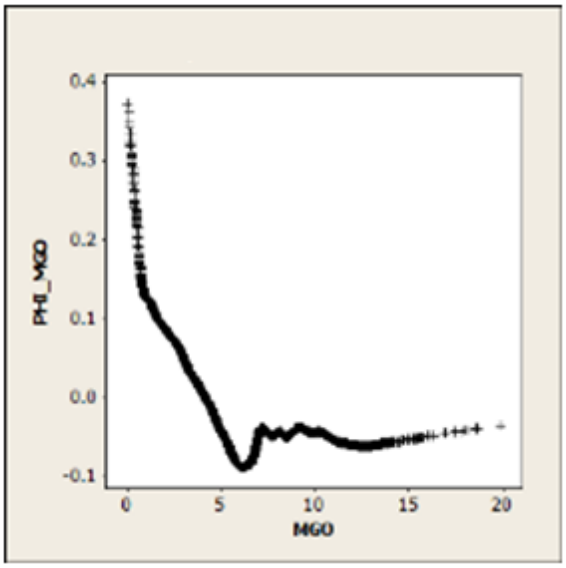

THETA_RMTOT
RMTOT

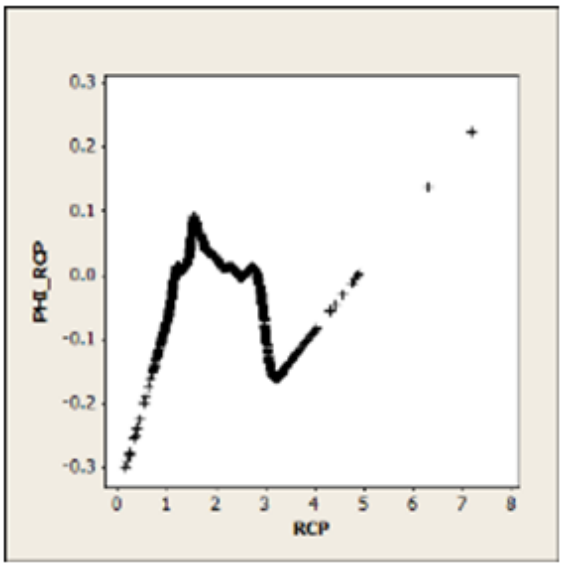

THETA_RMTOT
SUM_PHI

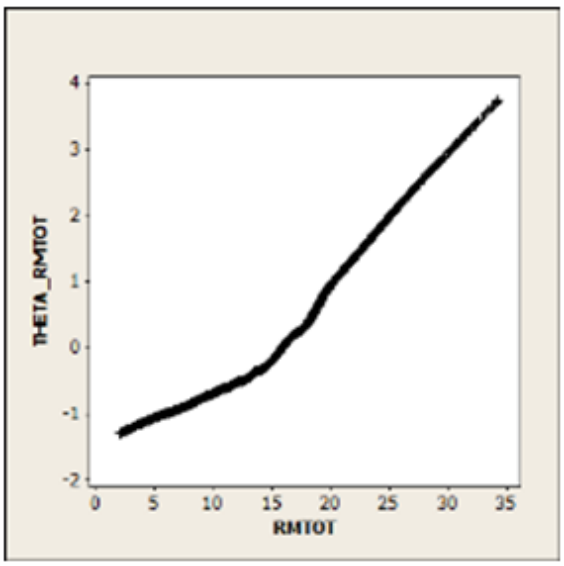
- Cenário J9:

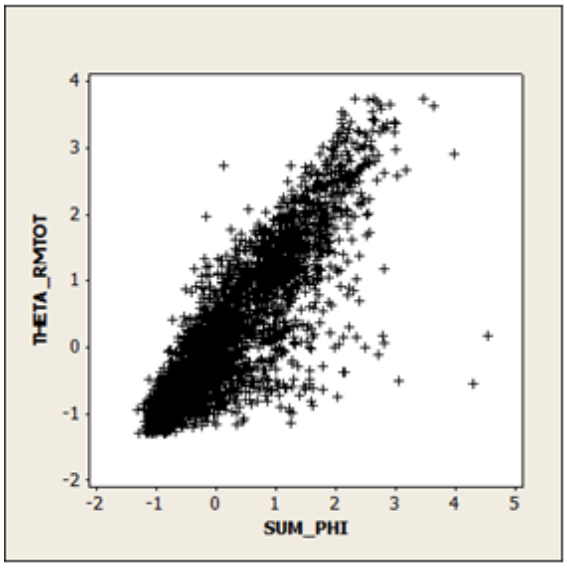

PHI_P2O5AP
P2O5AP


PHI_SIO2
SIO2

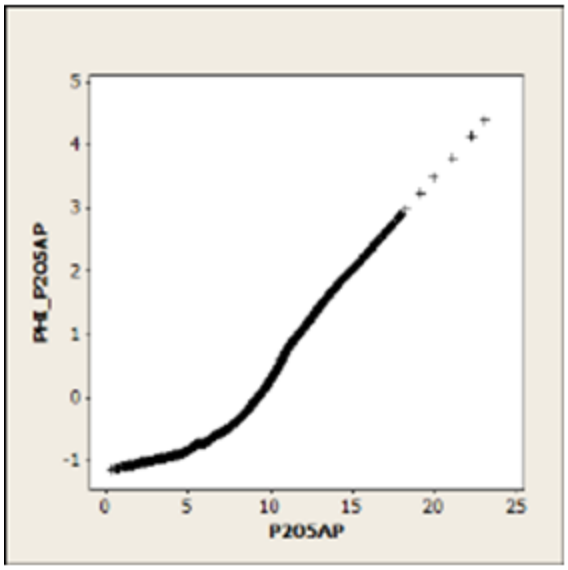

PHI_MGO
MGO

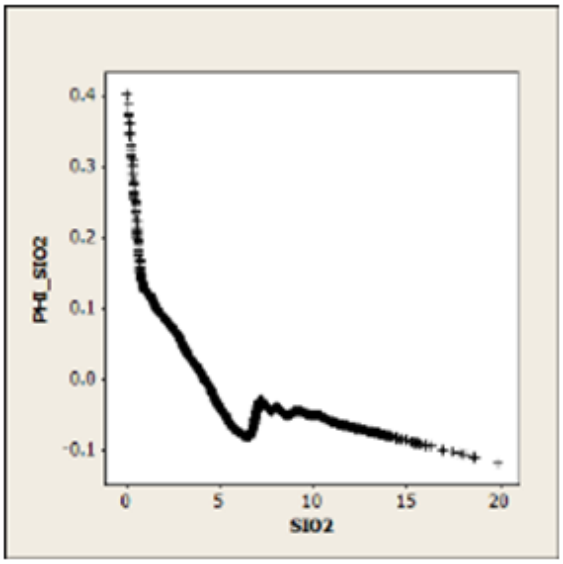

PHI_RCP
RCP

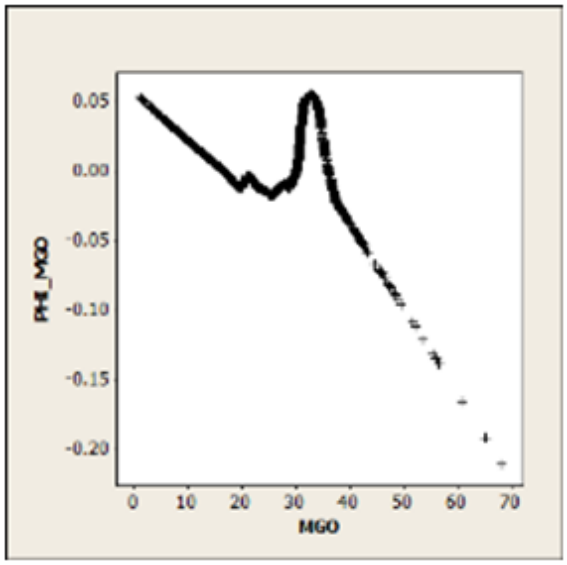

THETA_RMTOT
RMTOT

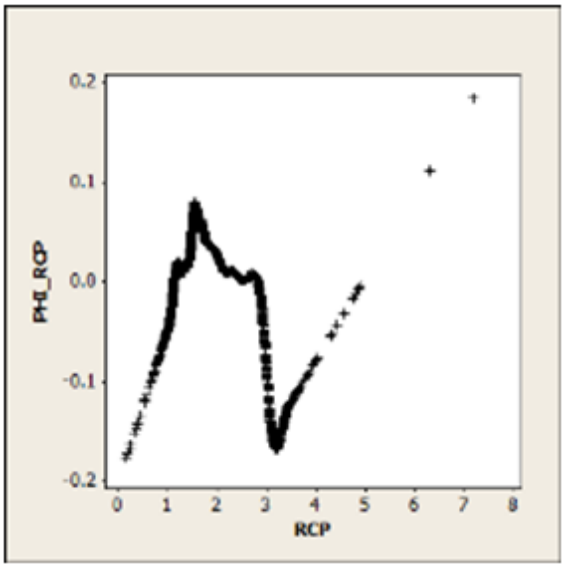

THETA_RMTOT
SUM_PHI

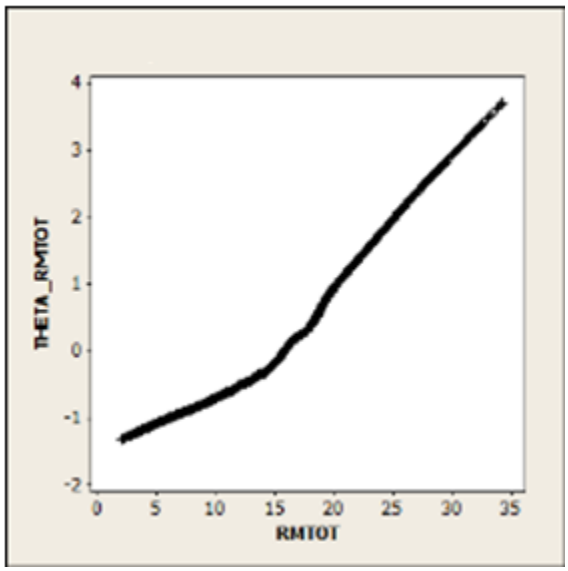
- Cenário J10:

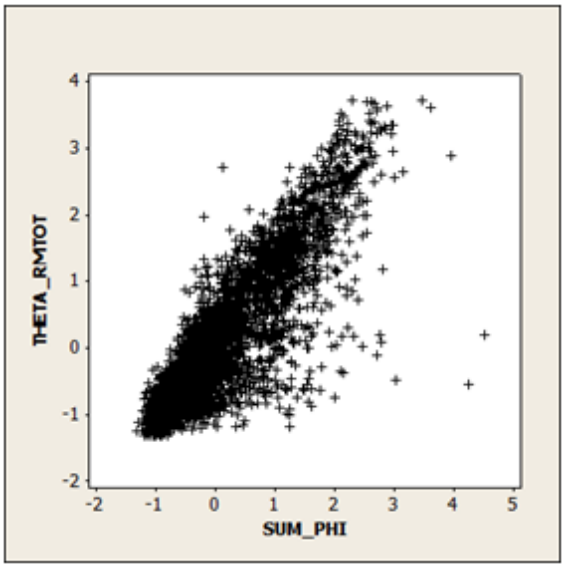

PHI_P2O5AP
P2O5AP


PHI_SIO2
SIO2

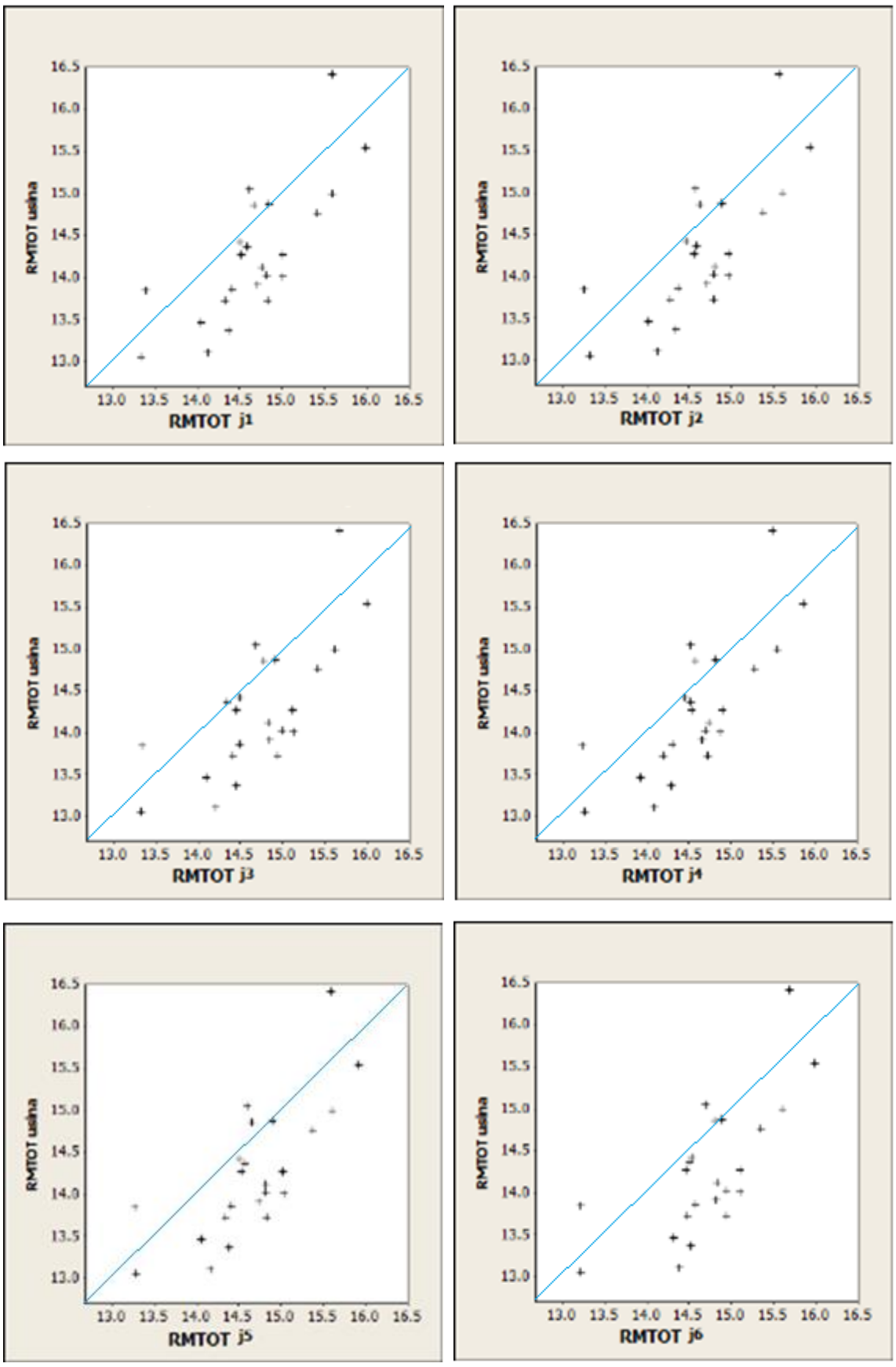

PHI_MGO
MGO

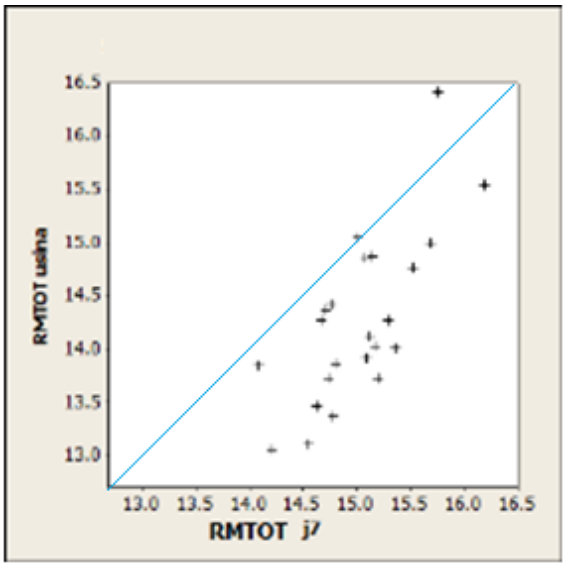

PHI_RCP
RCP

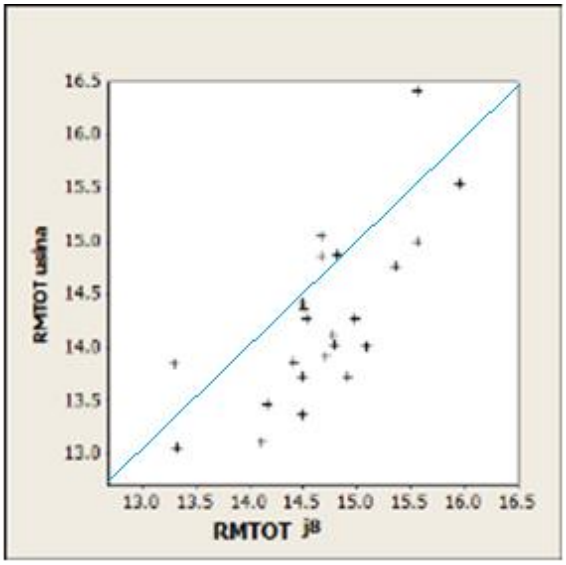

THETA_RMTOT
RMTOT

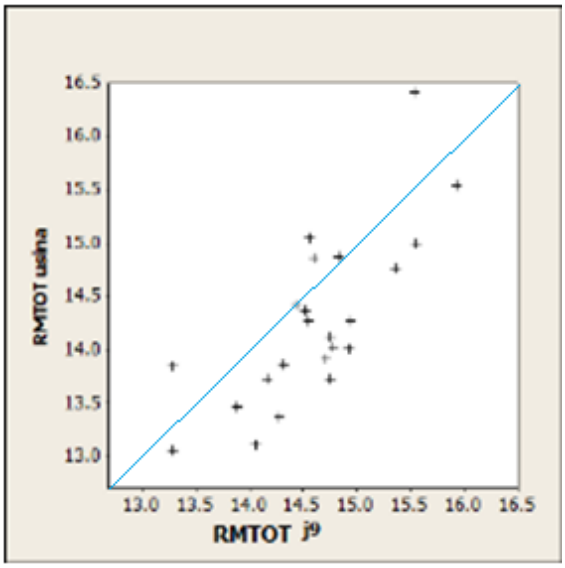

THETA_RMTOT
SUM_PHI

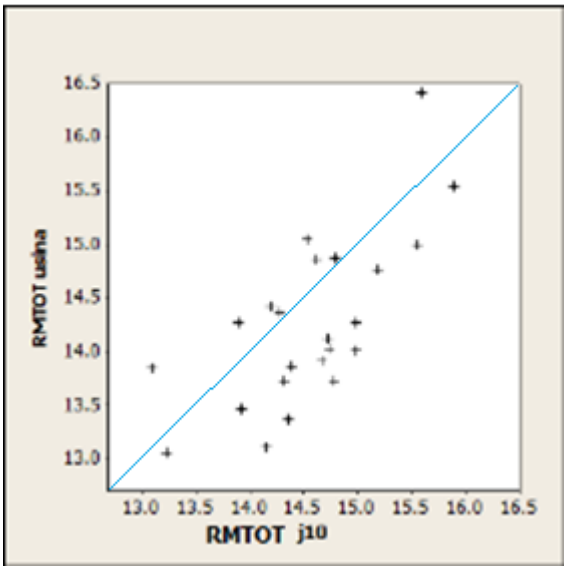
ANEXO 2: GRÁFICOS DE DISPERSÃO DO RMTOT USINA *VS* RMTOT DOS MODELOS



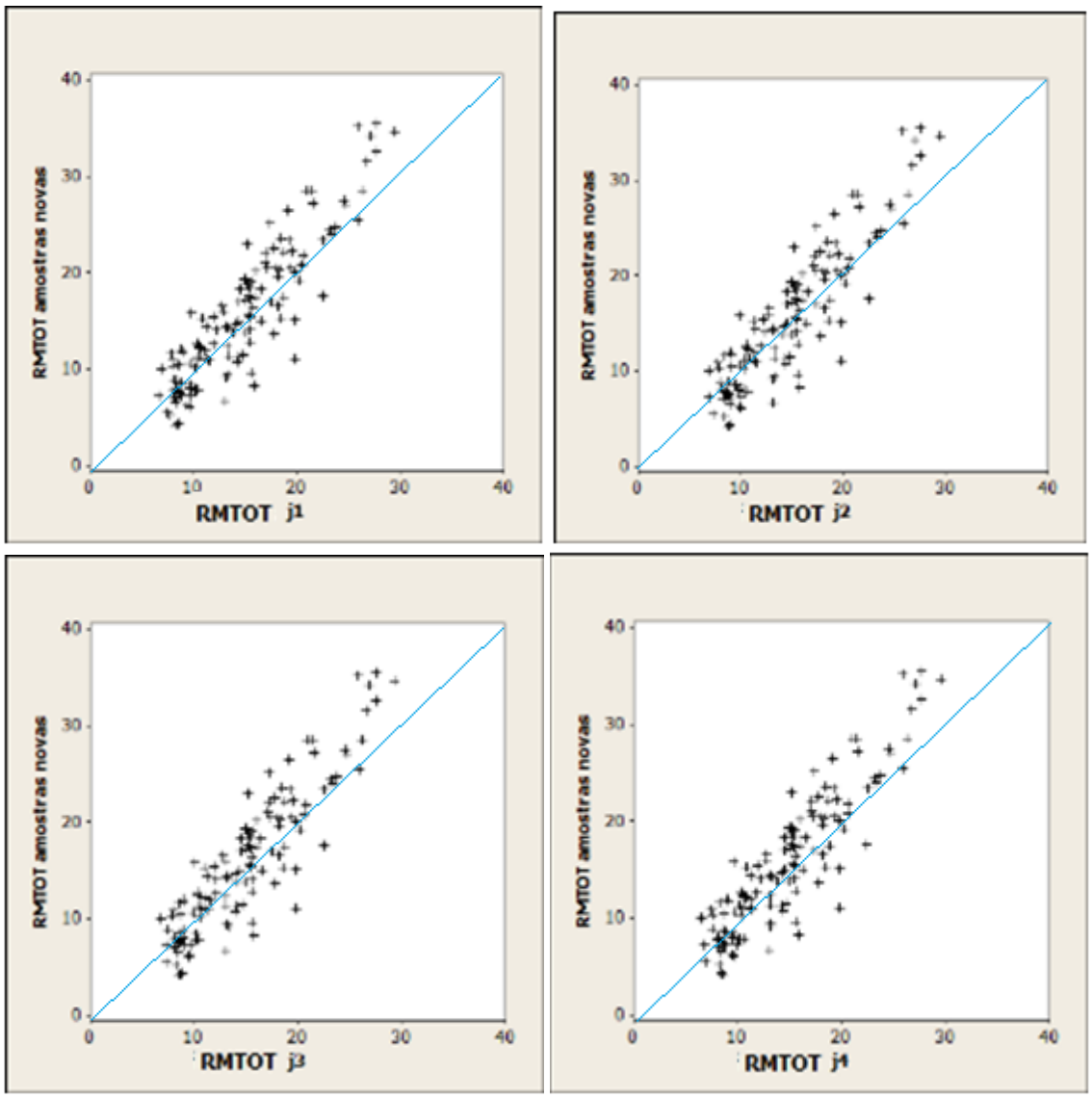

RMTOT usiha
16.5
16.0
15.5
15.0
14.5
14.0
13.5
13.0
13.0 13.5 14.0 14.5 15.0 15.5 16.0 16.5
RMTOT j7

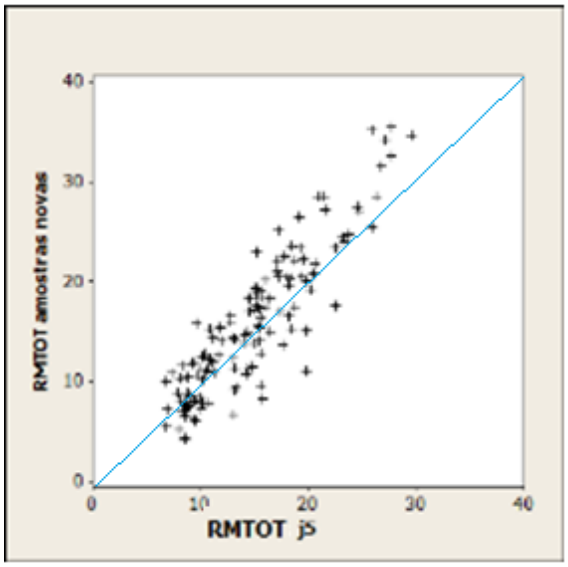

RMTOT usiha
16.5
16.0
15.5
15.0
14.5
14.0
13.5
13.0
13.0 13.5 14.0 14.5 15.0 15.5 16.0 16.5
RMTOT j8

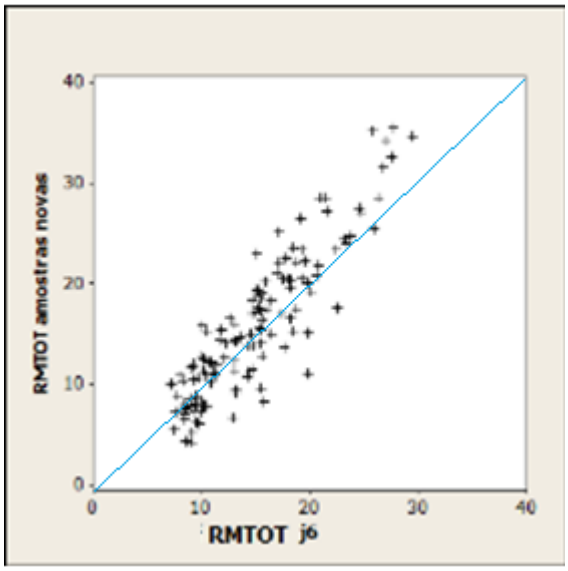

RMTOT usiha
16.5
16.0
15.5
15.0
14.5
14.0
13.5
13.0
13.0 13.5 14.0 14.5 15.0 15.5 16.0 16.5
RMTOT j9

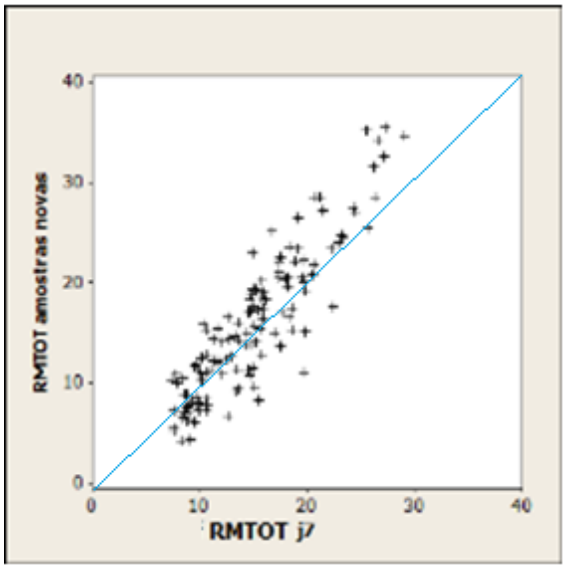

RMTOT usiha
16.5
16.0
15.5
15.0
14.5
14.0
13.5
13.0
13.0 13.5 14.0 14.5 15.0 15.5 16.0 16.5
RMTOT j10

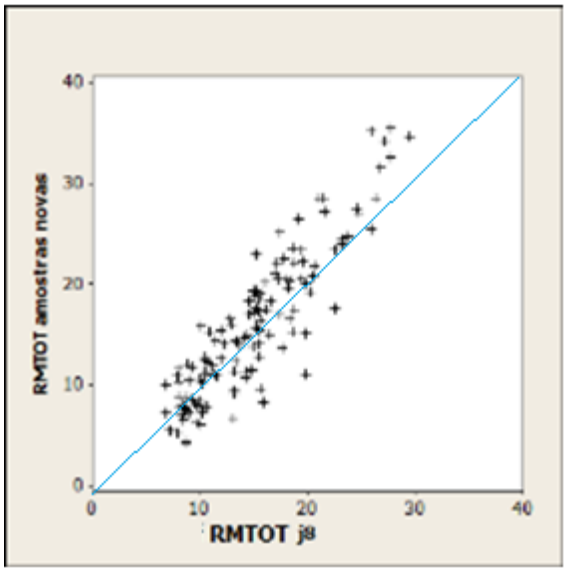
ANEXO 3: GRÁFICOS DE DISPERSÃO DOS RMTOT OBTIDOS PELOS MODELOS *VS* RMTOT COM AMOSTRAS NOVAS NA PLANTA PILOTO

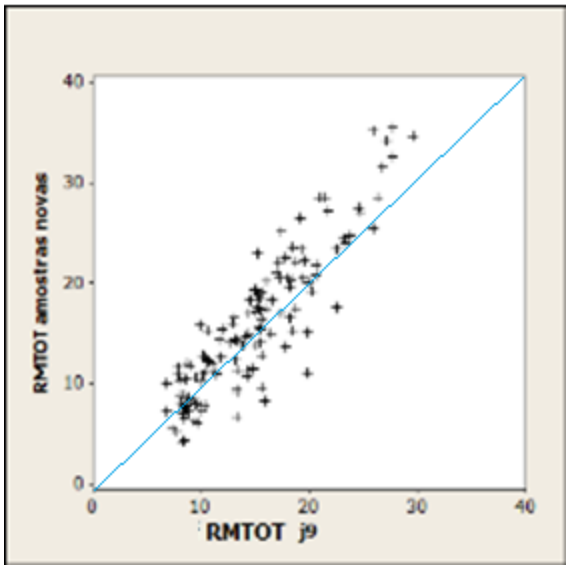


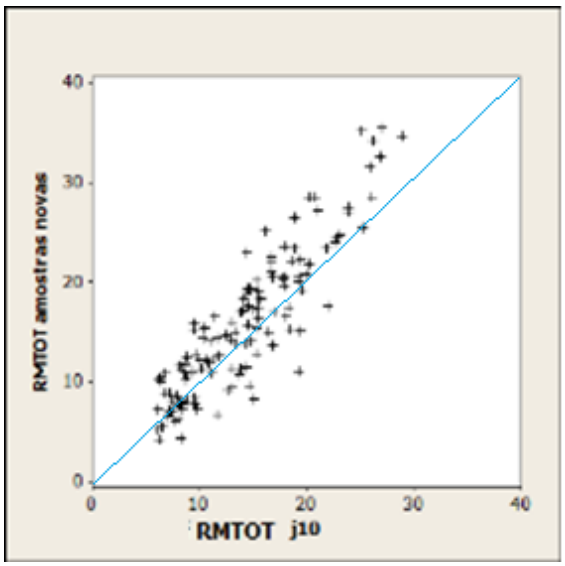

RMTOT amostras novas
RMTOT j5
0
10
20
30
40


RMTOT amostras novas
RMTOT j6
0
10
20
30
40


RMTOT amostras novas
RMTOT j7
0
10
20
30
40


RMTOT amostras novas
RMTOT j8
0
10
20
30
40


RMTOT amostras novas
RMTOT j9
0
10
20
30
40


RMTOT amostras novas
RMTOT j10
0
10
20
30
40