

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Instituto de Ciências Exatas

Departamento de Estatística

Douglas Rezende Mendonça

**CONSTRUÇÃO DE UM MODELO PREDITIVO PARA DETERMINAÇÃO DO TML
EM CARGAS DE FINOS DE MINÉRIO DE FERRO**

Belo Horizonte

2023

Douglas Rezende Mendonça

**CONSTRUÇÃO DE UM MODELO PREDITIVO PARA DETERMINAÇÃO DO TML
EM CARGAS DE FINOS DE MINÉRIO DE FERRO**

Monografia de especialização apresentada ao Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Especialista em Estatística.

Orientadora: Prof.^a Dra. Sueli Aparecida Mingoti

Belo Horizonte

2023

2023, Douglas Rezende Mendonça.
Todos os direitos reservados.

Mendonça, Douglas Rezende.

M539c Construção de um modelo preditivo para determinação do
TML em cargas de finos de minério de ferro [recurso eletrônico]
/ Douglas Rezende Mendonça. — 2023.
1 recurso online (44 f. il, color.): pdf.

Orientadora: Sueli Aparecida Mingoti .
Monografia (especialização) - Universidade Federal de Minas
Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: 43-44

1. Estatística. 2. Minério de ferro. 3. Análise de regressão. 4.
Transporte marítimo I. Mingoti, Sueli Aparecida. II.
Universidade Federal de Minas Gerais, Instituto de Ciências
Exatas, Departamento de Estatística. III. Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6/1510
Universidade Federal de Minas Gerais – ICEX




Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG


E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 292ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE DOUGLAS REZENDE MENDONÇA.

Aos vinte e dois dias do mês de maio de 2023, às 08:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **Douglas Rezende Mendonça**, intitulado: “Construção de um modelo preditivo para determinação do TML em cargas de finos de minério de ferro”, como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, a Presidente da Comissão, Professora Sueli Aparecida Mingoti – Orientadora, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje, por unanimidade. O resultado final foi comunicado publicamente ao candidato pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 22 de maio de 2023.

Documento assinado digitalmente
 SUELI APARECIDA MINGOTI
Data: 23/05/2023 17:26:03-0300
Verifique em <https://validar.iti.gov.br>

Prof.^a Sueli Aparecida Mingoti (Orientadora)
Departamento de Estatística / ICEX / UFMG

Documento assinado digitalmente
 ELA MERCEDES MEDRANO DE TOSCANO
Data: 24/05/2023 17:07:55-0300
Verifique em <https://validar.iti.gov.br>

Prof.^a Ela Mercedes Medrano de Toscano
Departamento de Estatística / ICEX / UFMG

Roberto da Costa
Quinino:80871291720
Assinado de forma digital por
Roberto da Costa
Quinino:80871291720
Dados: 2023.05.22 16:44:02 -03'00'

Prof. Roberto da Costa Quinino
Departamento de Estatística / ICEX / UFMG



Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
P Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

DECLARAÇÃO DE CUMPRIMENTO DE REQUISITOS PARA CONCLUSÃO DO CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA.

Declaro para os devidos fins que Douglas Rezende Mendonça, número de registro 2020680054, cumpriu todos os requisitos necessários para conclusão do curso de Especialização em Estatística e que entregou para sua orientadora, a professora Sueli Aparecida Mingoti, o trabalho, que aprovou a versão final. O trabalho foi apresentado no dia 22 de maio de 2023 com o título “Construção de um modelo preditivo para determinação do TML em cargas de finos de minério de ferro”.

Belo Horizonte, 10 de julho de 2023

Roberto da Costa
Quinino:8087129
1720

Assinado de forma digital
por Roberto da Costa
Quinino:80871291720
Dados: 2023.07.10 17:58:52
-03'00'

Prof. Roberto da Costa Quinino
Coordenador do curso de
Especialização em Estatística
Departamento de Estatística / UFMG

RESUMO

O minério de ferro é um dos mais relevantes produtos da economia nacional e significativo contribuidor na balança comercial brasileira, sendo o principal recurso mineral exportado pelo Brasil. Tendo quase a totalidade de seu volume transportado por vias marítimas, a cadeia produtora de mineração possui relevante preocupação com a umidade de suas cargas, cuja elevação descontrolada configura um risco à segurança das embarcações graneleiras envolvidas no transporte transoceânico. Controlar o teor de umidade nas cargas e seu limite máximo, conhecido como TML – limite de umidade transportável – é obrigação de todos os embarcadores de minério, segundo a IMO (*International Maritime Organization*), órgão regulador deste tipo de transporte. A obrigatoriedade da determinação do TML para as cargas de finos de minério insere na cadeia logística entre mina, ferrovia e porto, uma complexidade a mais durante a formação de suas cargas. Quando declarada inviável, uma nova proposta de carga exige a repetição do teste de TML, o que pode incorrer em atrasos nos embarques dos navios, causando prejuízos financeiros, tanto pela baixa na produtividade, quanto por multas e atrasos na dinâmica de afretamentos dos navios. Visto que o TML é um fator empírico, que deve ser previamente determinado para quase todas as cargas propostas, ter a possibilidade de conhecer antecipadamente um valor predito para o parâmetro se mostra bastante relevante para a cadeia produtiva, melhorando a capacidade de decisão e previsibilidade das operações. Neste contexto, este trabalho propõe um modelo estatístico preditivo para estimar o TML de cargas de finos de minério de ferro, partindo de parâmetros físico-químicos previamente analisados pelos laboratórios de produção. A aplicação da técnica de regressão linear múltipla, a partir de quatro variáveis independentes, produziu um modelo que se mostra capaz de predizer os valores do TML, com razoável assertividade, em um conjunto de dados teste. Esta abordagem, quando aplicada à rotina operacional, deverá configurar uma alternativa de planejamento que parte de dados já existentes, sem aumentar custos ou exigir acréscimo de recursos computacionais, contribuindo para o melhor andamento do sequenciamento de cargas ao longo da cadeia de produção de finos de minério de ferro.

Palavras-chave: Minério de Ferro. TML. Umidade. Regressão Linear. Transporte Marítimo.

ABSTRACT

Iron ore is one of the most relevant products of the national economy and a significant contributor in the Brazilian trade balance, being the main mineral resource exported by Brazil. Having almost all its volume transported by sea, the mining production chain has a relevant concern with the humidity of its cargoes, whose uncontrolled elevation constitutes a risk to the safety of bulk vessels involved in transoceanic transport. Controlling the moisture content in the loads and their maximum limit, known as TML – transportable moisture limit – is an obligation of all ore shippers, according to IMO (International Maritime Organization), the regulatory body of this type of transport. The mandatory determination of the TML for the loads of ore fines inserts in the logistics chain between mine, railroad and port, an additional complexity during the formation of their cargoes. When declared unfeasible, a new cargo proposal requires the repetition of the TML test, which may incur delays in vessel shipments, causing financial losses, both due to the decrease in productivity, as well as taxes and fines in chartering service. Since the TML is an empirical factor, which must be previously determined for almost all proposed loads, having the possibility of knowing in advance a predicted value for this parameter is very relevant to the production chain, improving the decision-making capacity and predictability of operations. In this context, this work proposes a predictive statistical model to estimate the TML of iron ore fines loads, starting from physicochemical parameters previously analyzed by production laboratories. The application of the multiple linear regression statistical method, with four independent variables, produced a model that is capable to predict values for TML with reasonable accuracy in a test dataset. This approach, when applied to the operational routine, should configure a planning alternative that starts from existing data, without increasing costs or requiring additional computational resources, contributing to a better progress in the sequencing of loads along the production chain of fines iron ore.

Keywords: Iron Ore. TML. Moisture. Linear Regression. Maritime Transport.

Lista de Figuras

Figura 1: Navio adernado por liquefação de carga granel.....	12
Figura 2: Histogramas e <i>Boxplot</i> para as variáveis FE, SIO ₂ , AL ₂ O ₃ , P e MN.....	17
Figura 3: Histogramas e <i>Boxplot</i> para as variáveis PPC, H ₂ O e TML.....	18
Figura 4: Histogramas e <i>Boxplot</i> para as variáveis de granulometria.....	20
Figura 5: Saída do Minitab para análise de resíduos do modelo.....	34
Figura 6: Gráficos de dispersão e de linhas do TML – Predito x Reais.....	38
Figura 7: Histograma e gráfico de linha dos desvios TML Predito – Laboratório.....	39
Figura 8: Avaliação prática do modelo – critérios de validação.....	40

Lista de Tabelas

Tabela 1: Lista de parâmetros disponíveis nos dados.....	15
Tabela 2: Estatísticas descritivas das variáveis de estudo	16
Tabela 3: Saída do Minitab acerca dos melhores subconjuntos.....	30
Tabela 4: Estimativas para os parâmetros do modelo de regressão proposto.....	31
Tabela 5: Saída do Minitab para coeficientes, valor-P e sumário geral do modelo	32
Tabela 6: ANOVA do Minitab para o modelo proposto.....	33
Tabela 7: Estatísticas descritivas das variáveis de estudo do conjunto teste.....	36
Tabela 8: Estatísticas descritivas para variável resposta – TML real x predito	37

Sumário

1. INTRODUÇÃO	10
1.1. Minério de ferro	10
1.2. Teor de Umidade e TML	10
1.3. Dinâmica de embarque de cargas de finos de minério de ferro	12
2. DESENVOLVIMENTO	15
2.1. Apresentação dos dados	15
2.2. Metodologia	21
2.2.1. Regressão linear múltipla.....	21
2.2.2. Procedimento de estimação dos parâmetros.....	22
2.2.3. Medidas estatísticas.....	23
2.2.3.1. Definições	23
2.2.3.2. Coeficientes de determinação R^2 , R^2 ajustado e R^2 predito	24
2.2.3.3. Coeficiente de Mallows (C_p).....	25
2.2.3.4. Critérios de Akaike Corrigido (AICc) e de informação Bayesiano (BIC).....	26
2.2.3.5. Erro médio (ME) e erro médio absoluto (MAE)	27
2.2.4. Testes de hipóteses para avaliação da regressão.....	27
2.3. Proposição do modelo e resultados	29
2.3.1. Análises de melhores subconjuntos	29
2.3.2. Construção do modelo.....	31
2.3.3. Avaliação da qualidade de ajuste do modelo.....	32
2.3.4. Avaliação de resíduos.....	34
3. DISCUSSÃO DOS RESULTADOS	36
3.1. Análise de dados do conjunto teste	36
3.2. Análise de eficácia e validação do modelo preditivo	37
4. CONCLUSÕES	42
REFERÊNCIAS	43

1. INTRODUÇÃO

1.1. Minério de ferro

O Brasil é um dos maiores produtores de minério de ferro do mundo. Presente em reservas minerais distribuídas ao longo do território brasileiro, principalmente nos estados de Minas Gerais e do Pará, esta *commodity* se consolida como uma das mais relevantes do mercado e um dos motores da economia nacional.

Tendo como seu destino majoritário o mercado asiático, o minério de ferro é largamente transportado por vias marítimas, utilizando navios graneleiros de grande porte na execução destes fretes. Os portos brasileiros, estando os mais importantes distribuídos nos estados do Maranhão, Espírito Santo e Rio de Janeiro, chegaram à marca 344,1 milhões de toneladas exportadas e a um faturamento na casa dos US\$ 28,9 bilhões em 2022 (IBRAM, 2023).

1.2. Teor de Umidade e TML

O teor de umidade (ou *moisture content*, MC), geralmente apresentado em percentual massa sobre massa (% m/m), é a quantidade mássica percentual de água presente em uma amostra quando comparada à sua massa total, esta também conhecida como massa úmida.

Controlar a umidade e trabalhar para reduzi-la é fundamental para o negócio da mineração, pois os produtos são vendidos na chamada “base seca”, a qual desconsidera a tonelagem de umidade presente na carga (FERREIRA, 2019). Considerando que os navios usam combustíveis fósseis, transportar mais água intrínseca à carga constitui um grande ofensor aos custos de exportação, visto que, além das grandes quantidades utilizadas, esta fonte de energia sofre influência direta das cotações de câmbio e do próprio petróleo.

O TML (*Transportable Moisture Limit*, ou limite de umidade transportável) é o parâmetro utilizado para definir o máximo valor que o teor de umidade (MC) de cargas

de granéis podem alcançar ao serem embarcadas em navios de transporte marítimo. Este parâmetro varia entre os diferentes minérios e até mesmo entre cargas de um mesmo produto (IBRAM, 2021).

Segundo FERREIRA et al. (2017), esta variação do TML provém das diversas características dos minérios que podem compor as cargas a serem analisadas. Fatores como a formação litológica, diferenciação mineralógica – como hematitas, goethitas e magnetitas – são determinantes para o patamar, por exemplo, de drenabilidade e potencial absorção de água pelo minério de ferro. Estas características, juntamente com outras condições de beneficiamento, distribuição granulométrica e degradação das partículas ao longo da cadeia logística, contribuirão para a caracterização da umidade do minério em sua máxima saturação permitida para transporte seguro – ou seja, o TML – por meio dos testes laboratoriais.

Neste ponto, cabe um destaque para distribuição granulométrica do material, em outras palavras, quão variados são os tamanhos das partículas presentes na carga. Materiais mais homogêneos, independentemente de serem compostos por partículas mais finas ou mais grosseiras, tendem a manter mais espaços vazios entre seus grãos, ao passo que, em minérios granulometricamente mais heterogêneos, as partículas mais finas tendem a ocupar o espaço entre as maiores, diminuindo assim o volume de vazios do material.

O volume de vazios é um parâmetro importante pois, durante o ensaio laboratorial, serão estes espaços vazios a acomodar as diferentes quantidades de água utilizadas nos testes, que levarão a diferentes saturações do minério e que serão a base empírica para determinar o TML daquele lote. O volume de vazios é, portanto, salvo em casos específicos de rara exceção, proporcional ao TML do material (FERREIRA, 2019).

Dentre os testes indicados para o minério de ferro, estão Proctor/Fagerberg, *Flow Table* ou *Penetration* (IMO, 2020). Cada tipo de minério citado na normativa de transporte marítimo é mais adequadamente descrito por algum dos métodos, sendo o Proctor/Fagerberg (PFC.70) e o Proctor/Fagerberg modificado (PFD.80) os mais adequados ao minério de ferro.

1.3. Dinâmica de embarque de cargas de finos de minério de ferro

Medir o TML da carga e controlar sua umidade é uma obrigação regulatória de todo embarcador de graneis. A IMO (*International Maritime Organization*), órgão internacional regulador de transportes marítimos, prevê em seu código (*IMSBC code*) todos os pontos a serem observados acerca destas operações.

Os chamados finos de minério de ferro, também conhecidos como *sinter feed*, são o principal insumo dos processos de sinterização nas siderúrgicas ao redor do mundo. Este tipo de material é classificado no código IMSBC como carga do grupo A, ou seja, que estão sujeitas à liquefação ao longo do transporte marítimo.

Fenômenos de liquefação podem ocorrer quando o teor de umidade da carga supera o TML, o que pode levar a deslocamentos de massa indesejados nos porões do navio e causam, eventualmente, prejuízos à estabilidade de navegação, danos estruturais à embarcação e, em casos extremos, colapso e posterior naufrágio do navio, conforme mostrado na Figura 1 (INTERCARGO, 2021).



Figura 1: Navio adernado por liquefação de carga granel

Para controle dos processos e cumprimentos dos procedimentos previstos pelos reguladores, são utilizados ensaios de laboratório para determinar umidade e TML das cargas. A técnica para determinar o TML baseia-se, resumidamente, em ensaios de compactação. A técnica de determinação de umidade, por outro lado, utiliza estufa com temperatura controlada para remoção da água, até que a massa testada seja

constante e o percentual de água possa ser conhecido pela diferença entre as massas seca e úmida.

Neste texto, são estudadas situações descritas para minério de ferro *sinter feed*, que tem seu TML determinado por intermédio do teste de Proctor/Fagerberg modificado, também chamado PFD.80. Nele, uma amostra representativa da carga é submetida a diferentes saturações de umidade, sendo o TML determinado na umidade equivalente a 80% de saturação naquele material.

Segundo a norma regulamentadora, entende-se que apenas no teste empírico, sem considerar demais fatores de influência externa, pode haver um desvio de 0,2 pontos percentuais (p.p.) para mais ou para menos. A repetibilidade e reprodutibilidade toleram, portanto, somadas, uma variação absoluta de 0,2 p.p.. Em outras palavras, se operadores diferentes, utilizando aparatos de testes diferentes para um mesmo material, obtiverem uma diferença entre dois ensaios menor do que 0,2 p.p. (em módulo), os testes são considerados válidos e coerentes entre si.

O ensaio de compactação (TML) demanda um tempo razoável para ser executado, pois cada carga precisa ser testada em diferentes graus de saturação de umidade, por meios dos quais se obtém uma curva para avaliação do teste. Nos momentos de melhor desempenho, é possível gerar resultados após 24h de trabalho. Considerando o cenário de produção ininterrupta da cadeia de mineração, este prazo pode ser extremamente nocivo para a produtividade dos portos, visto que a ausência do parâmetro inviabiliza o embarque, que deve ter seu TML necessariamente certificado antes do início do carregamento do navio. A situação ilustrada pode, portanto, provocar momentos de ociosidade nas linhas produtivas, causando prejuízos às operações enquanto aguardam os números para continuidade dos processos.

Com resultados em mãos, compara-se o TML com o teor de umidade previsto para a carga. Quando os ensaios apontam para uma condição de $MC < TML$, tudo está conforme esperado e os embarques podem seguir normalmente. Porém, se o cenário anterior não se concretizar, a carga é declarada inviável e outra carga deve ser proposta, o que exige novos ensaios no laboratório, atrasando ainda mais a retomada da produção (FERREIRA et al., 2017)

Neste contexto, dispor de uma ferramenta preditora do valor de TML pode ser interessante para evitar que ensaios sejam realizados em cargas não promissoras.

Por exemplo, se o preditor apontar para um valor de TML que está muito próximo ao valor de umidade do material, provavelmente haverá dificuldades no controle de umidade ao longo do embarque. Sabendo disso, os planejadores podem utilizar o preditor para programar cargas com mais rapidez e menor esforço, propondo planos potencialmente mais viáveis e com maiores chances de sucesso desde a concepção da carga.

Para construção de um potencial modelo preditor, estão disponíveis parâmetros físico-químicos do minério, que são analisados previamente à medida em que se realizam os embarques ferroviários rumo aos portos. São teores de ferro, sílica, fósforo, alumina, perda ao fogo (LOI/PPC), umidade, distribuição granulométricas, entre outros. Embora se saiba que o TML é influenciado por outros parâmetros e informações não disponíveis nos dados, como as já citadas formações litológicas e mineralógicas, entende-se os parâmetros disponíveis compõem informações suficientes para prever o valor do TML de determinada carga com assertividade razoável, possibilitando a criação de um modelo viável para utilização no cotidiano de planejamento.

Este texto tem como objetivo, portanto, propor um modelo preditivo para simulação do TML de cargas de finos de minério de ferro, a partir dos parâmetros físico-químicos disponíveis, previamente analisados e já inseridos no cotidiano dos laboratórios de produção, de modo a facilitar o processo decisório durante a formação das cargas e posterior embarque, sem necessidade de alocação de novos recursos à rotina das operações, sejam eles tecnológicos ou financeiros.

2. DESENVOLVIMENTO

2.1. Apresentação dos dados

Para desenvolvimento do trabalho proposto neste texto, estão disponíveis dados de 112 embarques realizados ao longo do ano de 2020. Outros 72 embarques, datados do primeiro semestre de 2021, serão utilizados posteriormente como conjunto de dados teste. Todas as linhas de dados aqui representadas contém informações completas para cada um dos parâmetros dispostos na Tabela 1, na qual é possível visualizar a categoria dos parâmetros, seus rótulos – ou abreviações – e a unidade.

Tabela 1: Lista de parâmetros disponíveis nos dados

Categoria	Rótulo	% em relação ao todo
Químicos	FE	Teor de Ferro
	SIO2	Teor de Sílica
	AL2O3	Teor de Alumina
	P	Teor de Fósforo
	MN	Teor de Manganês
	PPC	Perda ao fogo
	H2O	Umidade
Granulometria acumulada	GR+10	GR acima de 10mm
	GR+6,3	GR acima de 6,3mm
	GR-6,3	GR abaixo de 6,3mm
	GR+1	GR acima de 1mm
	GR-0,15	GR abaixo de 0,15mm
	GR-0,106	GR abaixo de 0,106mm
TML	TML	% umidade

Fonte: o autor.

Na Tabela 2, são observadas as estatísticas descritivas para cada um dos parâmetros definidos anteriormente. São 7 parâmetros químicos do minério, 6 malhas granulométricas de referência e a propriedade que será a variável resposta deste texto, o TML, já definido na introdução desta monografia.

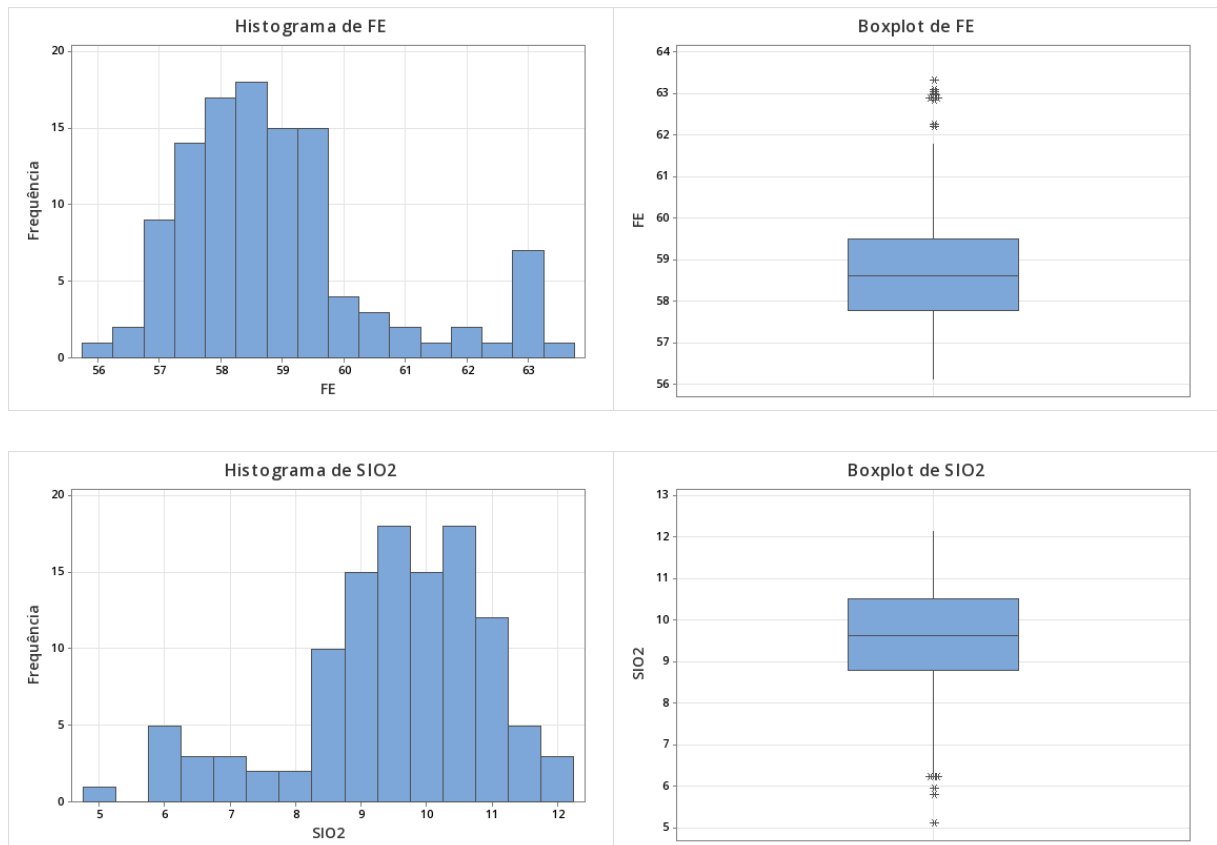
Sendo o ferro (Fe) o mineral de interesse e a sílica (SIO2) seu principal contaminante, verifica-se na Tabela 2 que estes são, respectivamente e conforme expectativas, os componentes com maiores médias no percentual mássico do material de estudo.

Tabela 2: Estatísticas descritivas das variáveis de estudo

Variável	Média	Desvio Padrão	Mínimo	1º quartil	Mediana	3º quartil	Máximo
FE	58,92	1,62	56,13	57,79	58,62	59,51	63,32
SIO2	9,49	1,46	5,11	8,81	9,64	10,51	12,15
AL2O3	2,01	0,32	1,01	1,89	2,06	2,23	2,63
P	0,066	0,015	0,043	0,054	0,061	0,080	0,095
MN	0,233	0,097	0,091	0,162	0,218	0,299	0,519
PPC	3,58	1,48	1,05	2,38	3,06	5,33	6,11
H2O	9,84	1,60	7,19	8,65	9,37	11,53	12,70
GR+10	5,94	1,88	1,19	5,05	5,97	6,82	11,30
GR+6,3	14,30	3,42	4,28	12,39	14,75	16,39	21,38
GR-6,3	85,70	3,42	78,62	83,61	85,26	87,61	95,72
GR+1	41,64	7,62	24,14	37,20	41,47	45,89	61,64
GR-0,15	37,22	9,52	16,27	29,91	38,53	44,12	54,32
GR-0,106	31,02	10,00	12,11	22,75	32,69	39,05	49,86
TML	11,63	1,58	9,05	10,26	11,18	13,45	14,34

Fonte: o autor.

Na sequência, na Figura 2, estão disponíveis os histogramas e o gráficos *Boxplot* para as 5 variáveis puramente químicas presentes nos dados: FE, SIO2, AL2O3, P e MN, deixando PPC e H2O para serem avaliadas juntas ao TML.



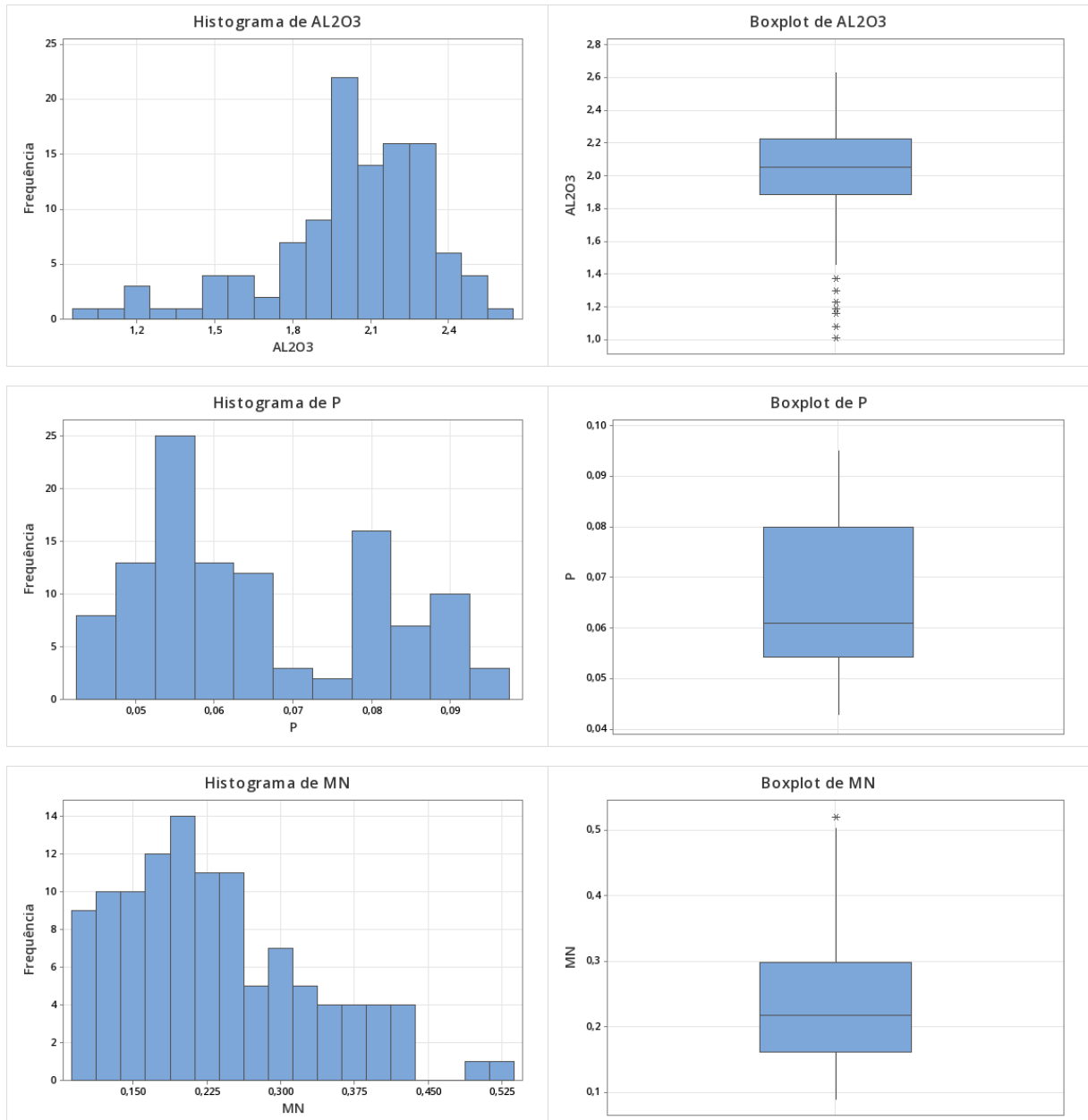


Figura 2: Histogramas e *Boxplot* para as variáveis FE, SiO₂, Al₂O₃, P e MN

Fe apresenta distribuição assimétrica à direita (positiva), com alguns *outliers* acima do quarto quartil. SiO₂, por sua vez, tem distribuição assimétrica à esquerda (negativa), com *outliers* abaixo do primeiro quartil. Este comportamento “complementar” entre os elementos é esperado por serem os principais componentes do minério de ferro, além de serem inversamente proporcionais na matriz mineral.

Al₂O₃ apresenta comportamento similar a SiO₂, com assimetria à esquerda e *outliers* abaixo do primeiro quartil. P e MN, por sua vez, apresentam assimetria à direita e poucos – ou nenhum – *outliers* em seus gráficos.

Na Figura 3, estão disponíveis os histogramas e os gráficos *Boxplot* para PPC, H2O e a variável resposta, TML. É sabido que o PPC e o TML tem relevante correlação (FERREIRA, 2019), enquanto o controle de umidade do embarque do navio é feito tendo o TML como seu limite superior. Desta forma, a semelhança no comportamento destes parâmetros – com assimetria levemente à direita e aparente ocorrência de duas modas nos gráficos – é coerente com a realidade empírica do material. A formação de dois patamares distintos de TML dentre os dados será comentado em seções posteriores.

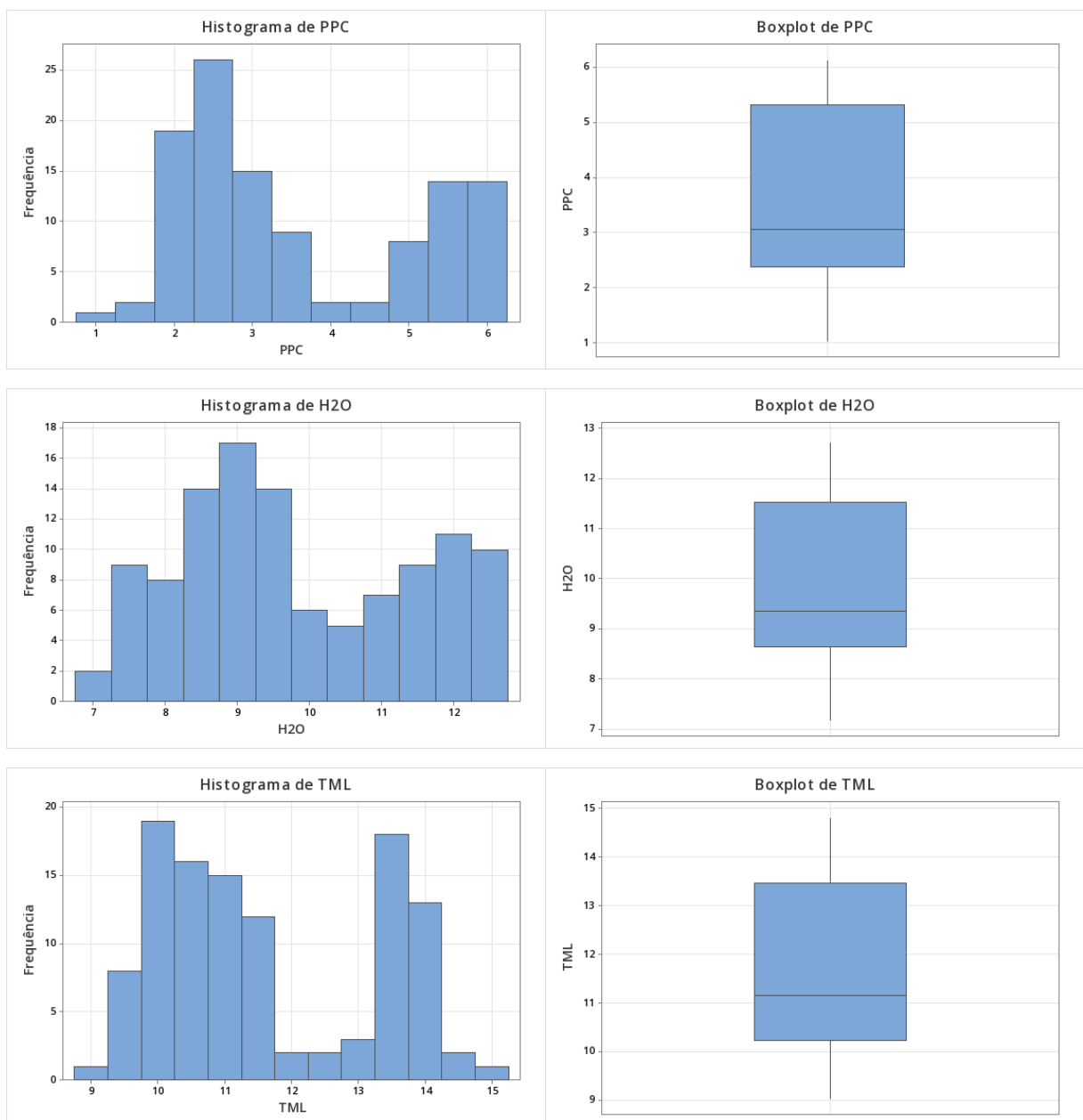


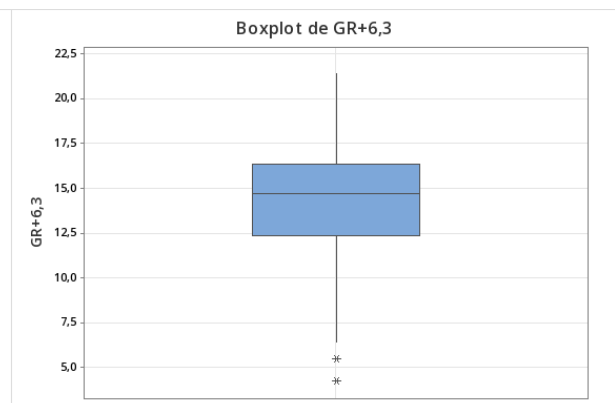
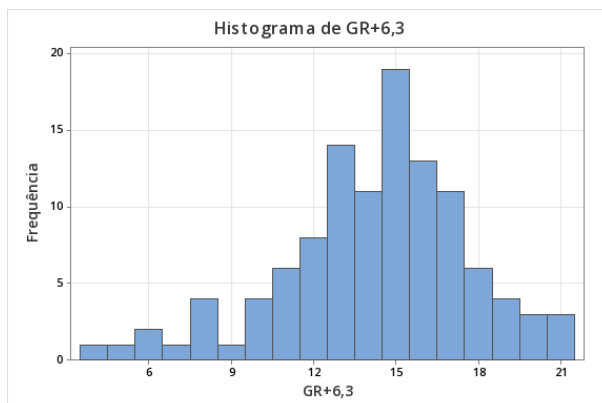
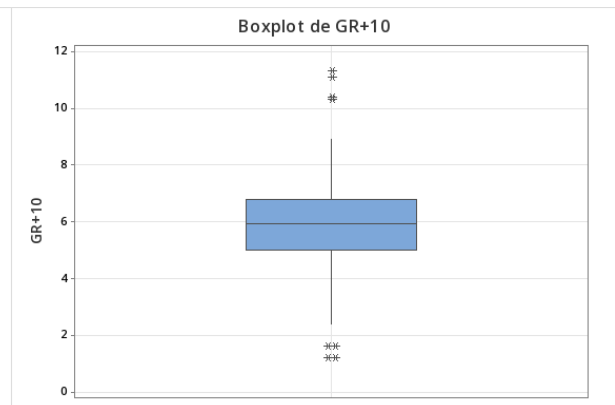
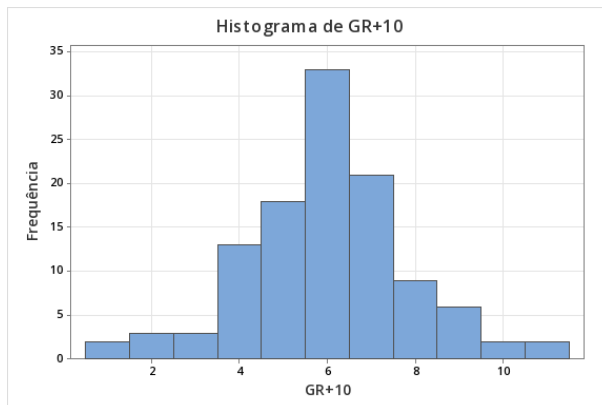
Figura 3: Histogramas e *Boxplot* para as variáveis PPC, H2O e TML

No caso da Figura 4, observa-se o comportamento gráfico das granulometrias presentes nos dados. Novamente, estão presentes alguns *outliers* para a maioria das peneiras.

GR+10 e GR+1 são as variáveis que apresentam simetria em suas distribuições, além da presença de *outliers* nos dois extremos de seus gráficos.

GR+6,3, GR-0,15 e GR-0,106 apresentam assimetria à esquerda e *outliers* apenas para GR+6,3, todos abaixo do primeiro quartil. Complementarmente, GR-6,3 apresenta assimetria à direita e *outliers* acima do quarto quartil.

A dispersão dos dados, principalmente nas peneiras mais grossas, aparenta ser menor que nas peneiras mais finas. Este é outro fato que faz sentido ao avaliar os dados de forma empírica. Normalmente, os produtos de finos de minério de ferro tem a maior parte de sua massa abaixo da granulometria de 6,3mm (-6,3mm), o que leva a uma maior concentração mássica nessa região, fazendo com que a maior variabilidade da distribuição ocorra nas peneiras abaixo desta faixa.



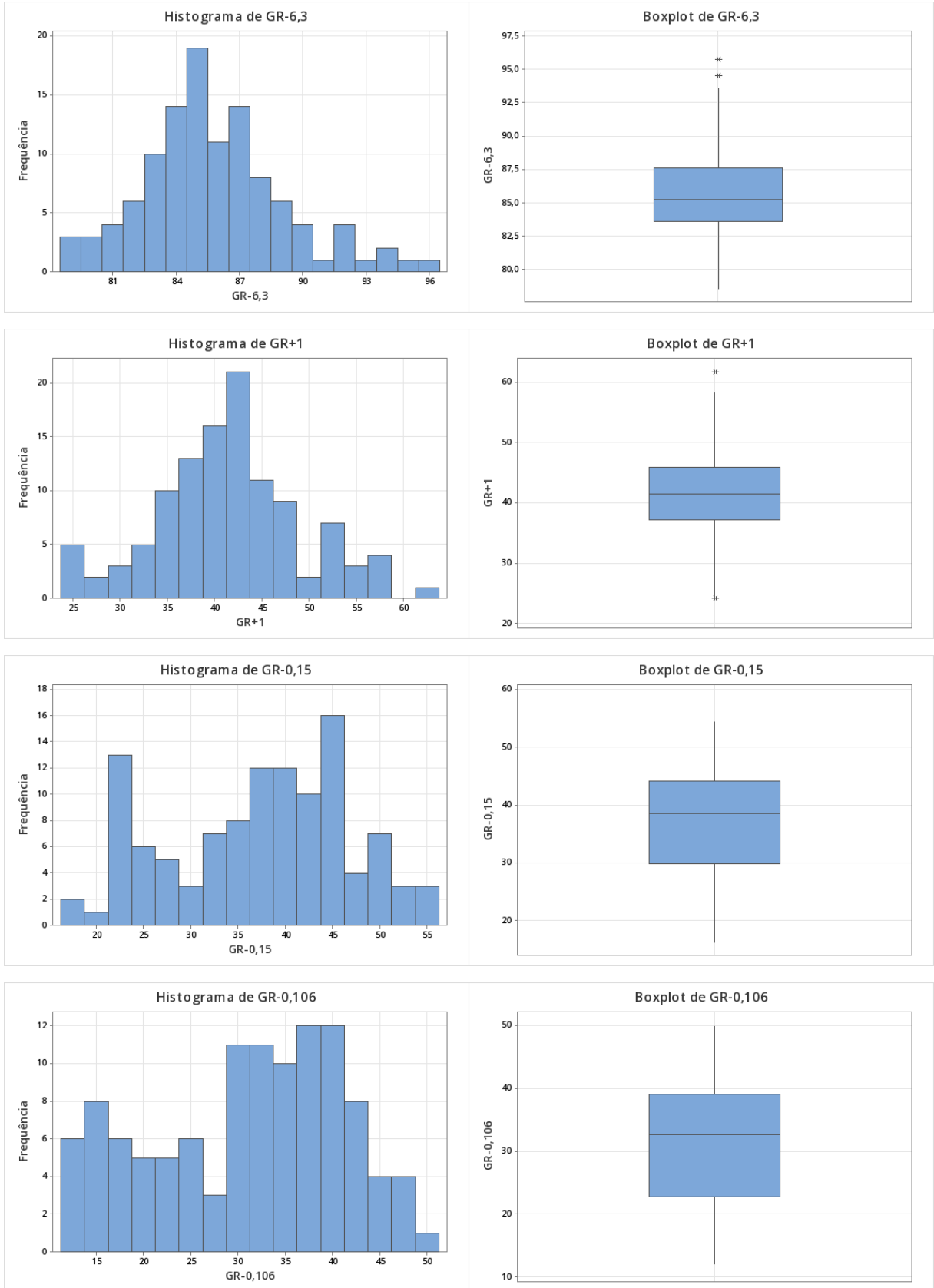


Figura 4: Histogramas e *Boxplot* para as variáveis de granulometria

2.2. Metodologia

2.2.1. Regressão linear múltipla

A regressão linear múltipla é um método estatístico que permite relacionar uma variável dependente (Y) com duas ou mais variáveis independentes $X_1, X_2 \dots X_k$. A definição matemática que descreve esta regressão é representada pela equação 1.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon \quad (1)$$

Nesta monografia, a variável dependente Y, também conhecida como variável resposta, é o TML. As variáveis independentes X_i , sendo $i = 1, 2, 3 \dots k$, por sua vez, são caracterizadas pelos elementos químicos, granulométricos e de umidade, conforme dados apresentados na Tabela 2 da seção 2.1. Já os coeficientes de regressão β_j , sendo $j = 0, 1, 2 \dots k$, são os parâmetros a serem estimados ao longo da construção de um modelo de regressão linear múltipla, sendo “k” o número de variáveis independentes do modelo.

Segundo MONTGOMERY & RUNGER (2012), o termo “linear” descrito no modelo é definido pelo fato de que a variável resposta, conforme equação 1 apresentada, é uma função linear dos parâmetros desconhecidos $\beta_0, \beta_1, \beta_2 \dots \beta_k$.

Ainda segundo os mesmos autores, o coeficiente β_0 representa a interseção do plano, em outras palavras, o valor médio de Y quando todas as variáveis independentes forem iguais a zero. Os coeficientes parciais de regressão, $\beta_1, \beta_2 \dots \beta_k$, são chamados desta forma pelo fato de medirem a variação esperada em Y por unidade de variação em sua respectiva variável independente, $X_1, X_2 \dots X_k$, quando as demais variáveis se mantiverem constantes.

O fator ϵ configura o termo de erro do modelo proposto pela equação 1, que se espera minimizar durante a modelagem e que deve, necessariamente, ser avaliado durante a validação do modelo. Supõe-se frequentemente os erros são não correlacionados, com média zero e variância σ^2 .

2.2.2. Procedimento de estimação dos parâmetros

O objetivo inicial na construção do modelo é, portanto, estimar os valores dos coeficientes de regressão (β_j) que melhor se ajustem aos dados observados. Esta estimação é realizada por meio do método de mínimos quadrados, em relação aos coeficientes $\beta_0, \beta_1, \beta_2 \dots \beta_k$, partindo das observações disponíveis nos dados e buscando minimizar a função L , mostrada na equação 2, adaptada de MONTGOMERY & RUNGER (2012).

$$L = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2 \quad (2)$$

O desenvolvimento da equação 2 é realizado com o objetivo de minimizar L . Para tal tarefa, deriva-se a função L em relação aos parâmetros β , como descrito nas equações 3 e 4, mostradas a seguir, também adaptadas de MONTGOMERY & RUNGER (2012).

$$\frac{\partial L}{\partial \beta_0} \Big|_{\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_k} = -2 \sum_{i=1}^n (y_i - \widehat{\beta}_0 - \sum_{j=1}^k \widehat{\beta}_j x_{ij}) = 0 \quad (3)$$

$$\frac{\partial L}{\partial \beta_j} \Big|_{\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_k} = -2 \sum_{i=1}^n (y_i - \widehat{\beta}_0 - \sum_{j=1}^k \widehat{\beta}_j x_{ij}) = 0 \quad (4)$$

A solução das derivadas leva a um conjunto de equações normais que estão descritas na obra de MONTGOMERY & RUNGER (2012). Após resolução das equações normais, são estimados os valores numéricos dos coeficientes de regressão, $\beta_0, \beta_1, \beta_2 \dots \beta_k$, que constituem o modelo da regressão linear múltipla.

Diversos *softwares* estatísticos são capazes de estimar os parâmetros da regressão a partir dos dados disponíveis, utilizando o método de mínimos quadrados descrito, sem que o usuário precise realizar cálculos separadamente.

No momento da avaliação da qualidade de ajuste da regressão linear múltipla, será importante considerar o comportamento dos resíduos do modelo proposto. Espera-se que eles apresentem comportamento normal, o que poderá ser confirmado por meio de análises gráficas, por exemplo, com uso de histogramas e gráficos de probabilidade, além de testes de aderência à normalidade, como os de Anderson-Darling e Shapiro-Wilk. Plotar os resíduos *versus* os valores observados da variável resposta também é útil, pois a existência de determinados padrões neste gráfico aponta para anomalias, que podem estar relacionadas à variância dos resíduos, levando a uma inadequação potencial do modelo.

Além dos resíduos, existem outros coeficientes e parâmetros a serem avaliados após a estimação dos parâmetros do modelo. Dentre eles, estão o coeficiente de determinação (R^2) e coeficiente de determinação ajustado (R^2 adj), estatística *Cp de Mallows* e dois tipos de medidas de erros, ME e MAE, descritos na seção 2.2.3 desta monografia.

2.2.3. Medidas estatísticas

2.2.3.1. Definições

Antes da apresentação das medidas, faz-se necessária a introdução de algumas definições, como as dos termos SQ_E , SQ_T , “n”, “p” e “k”.

A soma dos quadrados dos resíduos do modelo, SQ_E , é uma medida de variação dos resíduos em uma análise de regressão. Ela representa a soma dos quadrados das diferenças entre os valores observados da variável dependente e os valores previstos pela regressão linear múltipla, ou seja, a quantidade de variação na variável dependente que não foi explicada pelas variáveis independentes do modelo.

Paralelamente, a soma dos quadrados totais, SQ_T , é uma medida de variação total da variável dependente. Representa a soma dos quadrados das diferenças entre os valores observados da variável dependente e a média da variável dependente, isto é, a quantidade de variação total na variável dependente.

Observam-se as definições de SQ_E e SQ_T , respectivamente, nas equações 5 e 6.

$$SQ_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5)$$

$$SQ_T = \sum_{i=1}^n (y_i - \bar{y}_i)^2 \quad (6)$$

Por sua vez, “n” configura o número de observações disponíveis, o tamanho da amostra. O termo “k”, previamente apresentado na seção 2.2.1, representa o número de variáveis independentes no modelo, enquanto “p” representa o número de coeficientes estimados neste mesmo modelo. Entende-se, portanto, que $p = k + 1$, pois o termo β_0 também é contabilizado em “p”. Por último, nota-se a subtração “n – p”, que fornece os graus de liberdade dos resíduos do modelo.

2.2.3.2. Coeficientes de determinação R^2 , R^2 ajustado e R^2 predito

O Coeficiente de Determinação (R^2) é uma medida estatística que indica a quantidade da variabilidade total presente nos dados que é explicada ou considerada por um modelo de regressão ajustado. Esta medida varia entre 0 e 1, sendo mais favorável à qualidade do ajuste do modelo quanto mais próximo à unidade.

Alguns cuidados importantes devem ser tomados em relação ao R^2 , pois a adição indiscriminada de variáveis a um modelo deverá aumentar este coeficiente, sem necessariamente enriquecer sua qualidade e podendo, inclusive, deixá-lo menos ajustado do que antes da adição. Valores próximos a 1, por si, não garantem que o modelo seja adequado, nem garantem que o modelo forneça observações futuras de forma exata. Portanto, para regressão linear múltipla, a utilização desta medida estatística pode ser pouco adequada e recomenda-se outra abordagem para avaliação do modelo.

Configura-se como alternativa o uso do coeficiente de determinação ajustado, ou R^2 ajustado (R^2 adj), mostrado na equação 7.

$$R_{adj}^2 = 1 - \frac{SQ_E/(n - p)}{SQ_T/(n - 1)} \quad (7)$$

Visto que o denominador da equação 7 e o termo “n – p” são constantes, entende-se que R^2 ajustado somente aumentará quando uma nova variável adicionada contribuir com a redução de SQ_E . É justamente este o benefício do uso de R^2 (adj) em detrimento de R^2 , pois sua definição garante que a adição indiscriminada de variáveis ao modelo não será realizada sem uma possível penalização na qualidade de ajuste.

O R^2 predito (R^2 pred) é calculado a partir de uma fórmula que retira uma observação dos dados, estima novamente parâmetros para o modelo – sem considerar a observação retirada – e calcula o valor predito para esta observação deixada à parte. Isto é repetido de forma iterativa para cada uma das observações, mostrando ao final se o modelo fornece uma boa previsão. Como nos demais coeficientes, R^2 predito é um índice que varia entre 0 e 1 (MINITAB, 2023).

Um valor de R^2 predito significativamente menor que R^2 pode sinalizar excesso de ajuste do modelo, ao que passo que, caso R^2 predito seja próximo de 1, o modelo se mostra potencialmente capaz de prever a variável resposta a partir dos valores das variáveis independentes, fato particularmente útil no contexto desta monografia.

2.2.3.3. Coeficiente de Mallows (Cp)

O Coeficiente de *Mallows* (Cp), descrito na equação 8, é uma medida estatística que auxilia na avaliação da qualidade de um modelo de regressão. Ele é calculado comparando os valores ajustados por cada modelo estudado e os valores dos dados reais. SQ_E , σ^2 e “p” são, portanto, diretamente dependentes do modelo estudado.

$$C_p = \frac{SQ_E(p)}{\sigma^2} - n + 2p \quad (8)$$

De forma mais específica, o *Cp de Mallows* informa quão bem os valores ajustados por um modelo se ajustam aos dados reais. Numa outra abordagem, pode-se dizer que ele compara o modelo proposto com algo similar ao “melhor modelo possível”, com melhor ajuste de dados e quantidade otimizada de parâmetros selecionados.

Um valor de C_p próximo a “ p ” indica que o modelo avaliado é tão preciso quanto o “modelo ideal”, isto é, o qual estimaria parâmetros que, teoricamente, levariam a valores ajustados iguais aos valores reais, enquanto valores mais distantes de “ p ” indicam que o modelo se torna gradativamente menos preciso à medida em que se distancia do valor de “ p ”.

Este coeficiente é útil para comparar modelos alternativos de regressão e escolher aquele que melhor se ajusta aos dados, embora não deva ser utilizado de forma isolada na tomada de decisão (MONTGOMERY et al., 2015).

2.2.3.4. Critérios de Akaike Corrigido (AICc) e de informação Bayesiano (BIC)

O critério de Akaike Corrigido (AICc) e o critério de Informação Bayesiano (BIC) são medidas da qualidade relativa de um modelo de regressão linear múltipla e podem ser úteis para comparação entre diferentes propostas de modelo, sendo a melhor opção de modelo o que apresentar menor valor numérico para estes critérios (BURNHAM & ANDERSON, 2004).

Ambos são baseados no princípio de máxima verossimilhança e adicionam uma penalização ao modelo à medida em que novas variáveis são adicionadas.

Matematicamente, são definidos pelas equações 9 e 10, apresentadas a seguir, nas quais L é o valor numérico que maximiza a função verossimilhança do modelo, K o número de parâmetros e n o tamanho da amostra.

$$AICc = -2 * \log(L) + 2K + \frac{2K(K + 1)}{n - K - 1} \quad (9)$$

$$BIC = -2 * \log(L) + K * \log(n) \quad (10)$$

2.2.3.5. Erro médio (ME) e erro médio absoluto (MAE)

O erro médio (ME) e o erro médio absoluto (MAE) são potenciais medidas para avaliação da qualidade de ajuste de um modelo de regressão e estão definidos nas equações 11 e 12. Ambos são calculados como a média das diferenças entre os valores reais da variável resposta e as estimativas feitas pelo modelo, porém, no caso do MAE, este cálculo é realizado com os valores absolutos destas diferenças. As definições matemáticas para estas medidas estão dispostas nas equações a seguir.

$$ME = \sum_{i=1}^n \frac{y_i - \hat{y}_i}{n} \quad (11)$$

$$MAE = \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{n} \quad (12)$$

A avaliação do modelo por meio do erro médio (ME) deve ser feita com cautela, pois ele considera no somatório tanto as diferenças positivas, quanto as negativas. Para evitar a superestimação da qualidade de ajuste do modelo, o erro médio absoluto (MAE) pode servir como complemento da análise, pois considera o valor das diferenças em módulo, auxiliando na análise da real adequação de ajuste do modelo e sua capacidade de uso em soluções de predição, por exemplo. Estas duas medidas de erros serão úteis na avaliação do modelo em relação ao conjunto de dados teste.

2.2.4. Testes de hipóteses para avaliação da regressão

Quanto ao teste de significância da regressão, é possível propor um teste de hipóteses para avaliar os parâmetros estimados, considerando que a hipótese nula (H_0) apontaria para a não existência de relação linear entre os regressores e a variável

resposta, enquanto a rejeição desta hipótese informaria que pelo menos um deles deve ter relação linear entre si, conforme proposto na hipótese (H_1), mostrada na equação 13, com j variando entre 0 e k .

$$\begin{aligned} H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0 \\ H_1: \beta_j \neq 0, \text{ para pelo menos um } j \text{ entre } 1 \text{ e } k \end{aligned} \quad (13)$$

Adicionalmente, é realizada a avaliação da significância do modelo por meio de testes de hipóteses individuais para cada parâmetro. O teste realizado na equação 14 é o t de Student, que permite determinar se os coeficientes de regressão estimados têm um efeito significativo na variável de resposta. No contexto de regressão linear múltipla, o teste t é aplicado individualmente a cada coeficiente de regressão, avaliando se estes são estatisticamente iguais a zero, na hipótese nula (H_0), ou se, alternativamente (H_1), os coeficientes do modelo são estatisticamente diferentes de zero (KUTNER et al., 2018).

$$\begin{aligned} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0, \text{ para cada } j \text{ entre } 0 \text{ e } k \end{aligned} \quad (14)$$

Pelo teste da equação 14, avaliam-se as variáveis independentes, uma a uma, em relação à hipótese nula (H_0), que propõe que elas não possuem relação linear com a variável resposta.

A estatística t , utilizada na crítica do teste, pode ser visualizada na equação 15. Nela, $\hat{\beta}_j$ é o coeficiente individualmente avaliado, β_{H0} é $\hat{\beta}_j$ na hipótese nula – ou seja, igual a zero – e $SE_{\hat{\beta}_j}$ é o erro padrão do estimador.

$$t = \frac{(\hat{\beta}_j - \beta_{H0})}{SE_{\hat{\beta}_j}} = \frac{(\hat{\beta}_j - 0)}{SE_{\hat{\beta}_j}} \quad (15)$$

Outra avaliação importante se aplica aos resíduos, com uso do teste estatístico de Anderson-Darling, que verifica se a distribuição de uma determinada amostra de dados se assemelha a uma distribuição de probabilidade específica (KUTNER et al., 2018). No caso desta monografia, ele será aplicado aos resíduos do modelo obtido para avaliar a suposição de normalidade destes resíduos, conforme o teste de hipóteses da equação 16.

$$\begin{aligned} H_0: & \text{os resíduos atendem à normalidade} \\ H_1: & \text{os resíduos não atendem à normalidade} \end{aligned} \quad (16)$$

2.3. Proposição do modelo e resultados

2.3.1. Análises de melhores subconjuntos

O método de melhores subconjuntos, ou *Best Subsets*, consiste em uma técnica de seleção de variáveis para modelos de regressão linear múltipla. Ele envolve a construção de todos os subconjuntos possíveis das variáveis independentes e a posterior seleção do melhor subconjunto, levando em consideração um ou mais critérios a serem definidos pelo analista (KUTNER et al., 2018).

Como primeira abordagem, utilizando o software Minitab®, é possível observar quais das variáveis originais podem ser realmente úteis e adequadas para obtenção de um modelo adequadamente ajustado, a ponto de ser posteriormente utilizado como um modelo preditivo da variável resposta que, conforme objetivo, é o TML ($Y = \text{TML}$).

Numa análise preliminar dos dados e de acordo com a experiência empírica (FERREIRA, 2019), entende-se que o uso acumulado de dados de algumas frações granulométricas não influenciam no TML. Além disso, por serem parte de uma distribuição que se resume à somatória em 100% da massa por definição, observa-se uma forte correlação entre as componentes da análise granulométrica. Portanto, previamente à execução da análise de Melhores Subconjuntos do Minitab, foram retiradas as granulometrias +10mm, -6,3mm e -0,15mm.

A análise de subconjuntos produziu os resultados mostrados na Tabela 3. O valor para R^2 ajustado fica num patamar bastante razoável para todos os modelos apresentados, alcançando seu ponto ótimo nos modelos com 4 e 5 variáveis. Segundo MONTGOMERY & RUNGER (2012), este “ótimo” é um ponto interessante de partida para análise dos candidatos a melhor modelo de regressão.

Tabela 3: Saída do Minitab acerca dos melhores subconjuntos

Nro Vars	R-quad.	R2 (aj)	R2 (pred)	Cp	S	AICc	BIC	FE	SIO2	AL2 O3	P	MN	PPC	H2O	GR+ 6,3	GR+ 1	GR- 0,106
1	95	95	94,9	16	0,36	92,77	100,70						X				
1	83,5	83,3	82,9	305,4	0,66	227,63	235,56				X						
2	95,3	95,2	95	11,2	0,35	88,59	99,09		X				X				
2	95,3	95,2	95,1	11,4	0,35	88,82	99,32						X	X			
3	95,6	95,5	95,3	5,2	0,34	82,87	95,89		X				X		X		
3	95,6	95,5	95,3	5,8	0,34	83,52	96,55	X					X		X		
4	95,9	95,7	95,4	1,5	0,33	79,11	94,62		X				X	X		X	
4	95,9	95,7	95,5	1,5	0,33	79,14	94,65		X				X	X	X		
5	95,9	95,7	95,4	1,9	0,33	79,66	97,62		X				X	X	X	X	
5	95,9	95,7	95,4	2,5	0,33	80,29	98,25		X				X	X		X	X
6	95,9	95,7	95,3	3,3	0,33	81,27	101,62		X				X	X	X	X	X
6	95,9	95,7	95,3	3,6	0,33	81,68	102,03		X			X	X	X	X	X	X
7	96	95,7	95,2	5,1	0,33	83,44	106,15		X			X	X	X	X	X	X
7	95,9	95,7	95,2	5,2	0,33	83,55	106,25		X		X		X	X	X	X	X
8	96	95,6	95,1	7	0,34	85,78	110,79		X		X	X	X	X	X	X	X
8	96	95,6	95,1	7,1	0,34	85,84	110,85	X	X	X			X	X	X	X	X
9	96	95,6	95	9	0,34	88,23	115,49		X	X	X	X	X	X	X	X	X
9	96	95,6	94,9	9	0,34	88,24	115,50	X	X	X	X		X	X	X	X	X
10	96	95,6	94,8	11	0,33842	90,722	120,192	X	X	X	X	X	X	X	X	X	X

Modelos Possíveis

Modelo Escolhido

Fonte: o autor, a partir das análises dos dados originais feitas no Minitab.

Outra componente para avaliação é estatística C_p , ou coeficiente de *Mallows*. De acordo com MONTGOMERY & RUNGER (2012), este parâmetro auxilia na verificação de tendenciosidade do modelo proposto. Orienta-se que ele não se distancie de um patamar similar ao número de variáveis indicadas no subconjunto em questão. Novamente, há alguns modelos possíveis com valores favoráveis para C_p , destacando-se o intervalo entre 3 ou mais variáveis.

Adicionalmente, utiliza-se como análise complementar os critérios de Akaike Corrigido (AICc) e o Critério de Informação Bayesiano (BIC). Ambos são medidas que devem ser menores quanto melhor for o ajuste do modelo e estão minimizados nos modelos destacados na Tabela 3.

Como se sabe, não é interessante adicionar variáveis excessivamente ao modelo e os critérios aqui utilizados penalizam esta prática. Portanto, observando os conjuntos propostos, entende-se que o melhor deles, ao se considerar os valores R^2 ajustado, coeficiente *Cp de Mallows* e os critérios AICc e BIC, fica com as variáveis SIO2, PPC, H2O e mais uma peneira da granulometria, num modelo de 4 variáveis.

2.3.2. Construção do modelo

Na Tabela 3, apresentam-se duas alternativas de modelo com 4 variáveis e medidas de avaliação muito similares. Ambas diferem apenas na variável de granulometria a ser incluída, GR+1mm ou GR+6,3mm. Opta-se pela variável GR+1mm para compor o modelo de regressão, pois sabe-se que a quantidade de finos determinada nesta granulometria é um parâmetro crítico para preenchimento dos “espaços vazios” do minério, fator bastante influente nos resultados de TML (FERREIRA, 2019).

Decide-se, portanto, por um modelo de 4 variáveis, isto é, a quantidade “k” trabalhada nesta monografia assume valor igual a 4, com os parâmetros estimados na equação 17 e dispostos na Tabela 4. As críticas ao modelo proposto são realizadas com base nas recomendações previstas no texto de MONTGOMERY & RUNGER (2012).

$$TML = 6,941 + 0,0753 (SIO2) + 0,9814 (PPC) + 0,1132 (H2O) - 0,01574 (GR + 1) \quad (17)$$

Tabela 4: Estimativas para os parâmetros do modelo de regressão proposto

Variável Correspondente	Coeficientes Estimados	
Intercepto	$\widehat{\beta}_0$	6,941
SIO2	$\widehat{\beta}_1$	0,0753
PPC	$\widehat{\beta}_2$	0,9814
H2O	$\widehat{\beta}_3$	0,1132
GR+1	$\widehat{\beta}_4$	-0,01574

Fonte: o autor.

2.3.3. Avaliação da qualidade de ajuste do modelo

Para avaliação das estimativas dos parâmetros do modelo, respectivas significâncias estatísticas e multicolinearidade, tem-se na Tabela 5 os valores fornecidos pelo Minitab. Iniciando-se pela observação do fator de multicolinearidade (FIV ou VIF), nota-se um ponto de maior atenção nos valores para PPC e H2O, embora, neste caso, não seja um problema grave. A existência de multicolinearidade entre as variáveis explicativas pode prejudicar significativamente a estimação dos parâmetros β da regressão e, idealmente, os valores deste fator de avaliação do modelo deveriam ser iguais a 1 (MONTGOMERY & RUNGER, 2012).

Porém, ainda assim, considera-se que valores de VIF entre quatro e cinco, notados na Tabela 5, não devem gerar grandes preocupações em relação à qualidade da estimação dos parâmetros do modelo. Se seus valores estivessem mais próximos de 10 ou superassem este patamar, então haveria motivos para intervenção no modelo e possível retirada de variáveis correlacionadas (KUTNER et al., 2018).

Tabela 5: Saída do Minitab para coeficientes, valor-P e sumário geral do modelo

Termo	Coeficiente	Erro Padrão do Coeficiente	Valor-T	Valor-P	VIF
Constante	6,941	0,346	20,07	0,000	
SIO2	0,0753	0,0233	3,24	0,002	1,15
PPC	0,9814	0,0484	20,29	0,000	5,17
H2O	0,1132	0,0438	2,58	0,011	4,95
GR+1	-0,01574	0,00538	-2,93	0,004	1,68
S		R2	R2(adj)	R2(pred)	
0,332847		95,86%	95,70%	95,44%	

Fonte: o autor, a partir das análises dos dados originais feitas no Minitab.

Acerca do sumário da Tabela 5, avalia-se que o modelo está, *a priori*, satisfatório para descrever os dados originais. O parâmetro R^2 ajustado está em 95,70%, fato que indica elevada capacidade do modelo em explicar a variabilidade dos dados utilizados em sua construção.

Quanto ao teste de significância da regressão: a hipótese nula (H_0) apontaria para a não existência de relação linear entre os regressores e a variável resposta. Por outro

lado, a rejeição de H_0 leva à hipótese alternativa, que informa que pelo menos um deles teria relação linear, conforme proposto pelo teste (13) da seção 2.2.4.

Neste sentido, ao se avaliar a primeira linha da Tabela 6, que contém a análise de variância do modelo, entende-se que existe a relação linear aqui verificada, pois, pelo Valor-P = 0,000 correspondente, rejeita-se a hipótese nula (H_0) a um nível de significância $\alpha = 0,05$.

Tabela 6: ANOVA do Minitab para o modelo proposto

Fonte	GL	SQ (Aj.)	QM (Aj.)	Valor F	Valor-P
Regressão	4	274,231	68,5578	618,82	0,000
SIO2	1	1,162	1,1618	10,49	0,002
PPC	1	45,625	45,6246	411,82	0,000
H2O	1	0,739	0,7393	6,67	0,011
GR+1	1	0,949	0,9490	8,57	0,004
Erro	107	11,854	0,1108		
Total	111	286,086			

Fonte: o autor, a partir das análises dos dados originais feitas no Minitab.

Deve-se conduzir, individualmente, uma análise análoga para cada variável do modelo, avaliando as demais linhas da Tabela 6 para as variáveis independentes, em acordo com os testes de hipóteses (14) e (15) da seção 2.2.4 deste texto.

Para tal tarefa, aplica-se o teste de Student e a estatística t – ou seu valor-P correspondente – é avaliada em relação às hipóteses. Verificam-se, portanto, os valores-P, um a um, em relação à hipótese nula (H_0), que propõe que a variável independente correspondente do modelo não apresenta relação linear com a variável resposta.

Observando-se, portanto, as demais linhas da Tabela 5, nota-se que os valores-P mostrados (SIO2 = 0,002, PPC = 0,000, H2O = 0,011, GR+1 = 0,004), apontam para rejeição de H_0 , a um nível de significância $\alpha = 0,05$, para todas as variáveis independentes presentes no modelo, indicando assim a existência de relação linear entre cada uma delas e a variável resposta.

2.3.4. Avaliação de resíduos

Em relação aos resíduos do modelo, é possível visualizar suas características e comportamentos nos gráficos da Figura 5. Os gráficos à esquerda, de probabilidade normal e histograma, apontam para uma aparente normalidade na distribuição dos resíduos. A propósito, no primeiro gráfico, estão presentes os resultados do teste estatístico de Anderson-Darling (AD = 0,388 e Valor-P = 0,382), que permitem avaliar a hipótese de normalidade dos resíduos, também descrito na seção 2.2.4.

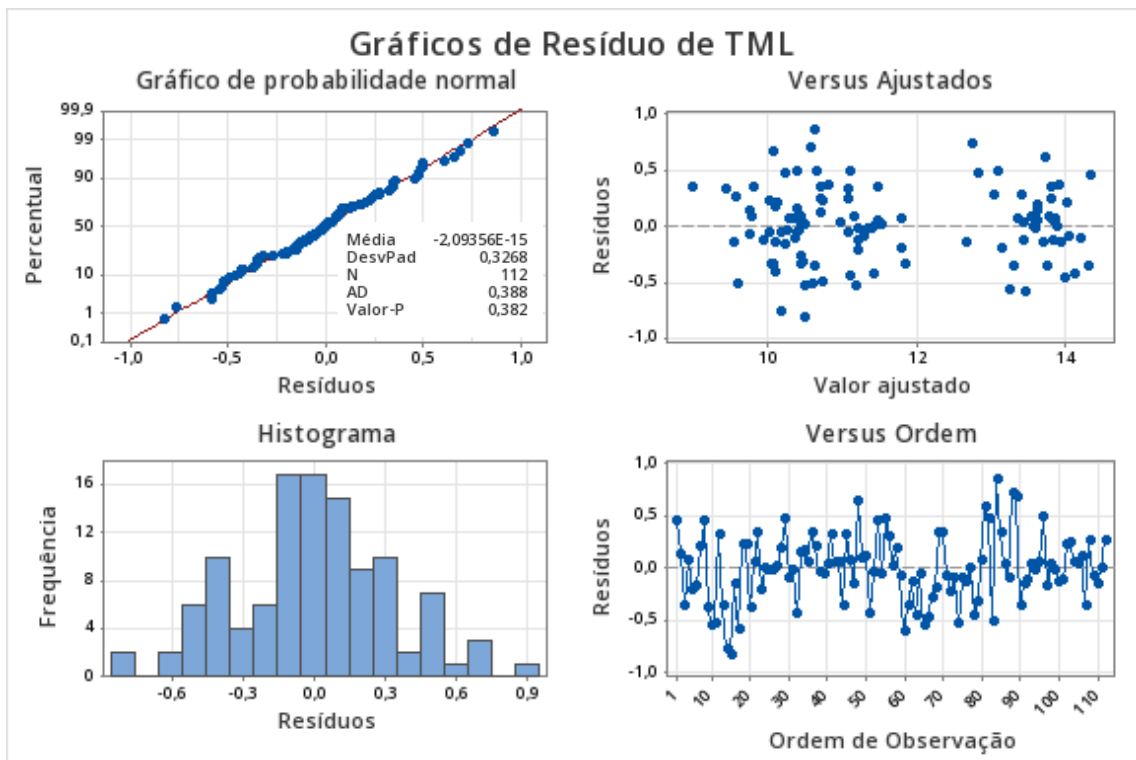


Figura 5: Saída do Minitab para análise de resíduos do modelo

Considerando o valor-P = 0,382 do teste de Anderson-Darling, e comparando-o com um nível de significância $\alpha = 0,05$, não é possível rejeitar a hipótese nula (H_0) de que os resíduos estão normalmente distribuídos.

Também, numa análise dos gráficos à direita na Figura 5, não são observados padrões nas plotagens dos resíduos, seja contra os valores ajustados pelo modelo ou mesmo plotado por ordem de observação.

No gráfico do canto superior direito, porém, dividem-se notadamente dois grupos. Estão localizados, aproximadamente, nos intervalos 9-12 e 13-15 das abscissas e são provenientes da distinção entre dois tipos de produto de finos de minério de ferro produzidos pela companhia. Embora sejam semelhantes em suas composições químicas, seus patamares de umidade e TML diferem entre si e esta diferença fica evidente no gráfico apresentado.

Levando em consideração o resultado do teste de hipóteses para normalidade (Anderson-Darling) e a avaliação visual dos gráficos fornecidos pelo Minitab, pode-se inferir que os resíduos do modelo possuem comportamento dentro das expectativas previamente introduzidas na seção 2.2, com comportamento normal e sem padrões anômalos nos gráficos. Dito de outra forma, não há inconsistências evidentes que desabonem o modelo proposto na tarefa de representar os dados originais, bem como em sua posterior aplicação no conjunto de dados para testes de eficiência na predição.

Em relação à variável resposta ($Y = \text{TML}$), cabe destacar o intervalo coberto pelos extremos deste parâmetro nos dados originais. Observa-se, conforme exposto na seção 2.1, que os valores mínimo e máximo para a variável são 9,05 e 14,79, respectivamente. Estes valores extremos serão importantes para avaliação da área de cobertura do modelo ao compará-lo com o intervalo observado no conjunto de teste.

3. DISCUSSÃO DOS RESULTADOS

3.1. Análise de dados do conjunto teste

Dispõe-se de um segundo conjunto de dados, medido ao longo do primeiro semestre de 2021. Suas respectivas estatísticas descritivas estão dispostas na Tabela 7, que representa as 72 linhas de dados que compõe o segundo conjunto, aqui definido como conjunto de teste.

Tabela 7: Estatísticas descritivas das variáveis de estudo do conjunto teste

Variável	Média	Desvio Padrão	Mínimo	1º quartil	Mediana	3º quartil	Máximo
FE	57,07	1,85	54,69	55,95	56,77	57,31	63,04
SIO2	11,42	1,74	6,46	10,81	11,63	12,58	15,95
AL2O3	1,95	0,30	1,04	1,85	1,98	2,11	2,47
P	0,075	0,012	0,054	0,064	0,078	0,086	0,095
MN	0,216	0,082	0,102	0,151	0,186	0,295	0,454
PPC	4,23	1,47	1,10	2,78	5,02	5,47	6,10
H2O	10,51	1,49	8,06	8,88	11,18	11,71	12,84
GR+10	6,15	1,71	1,66	5,39	6,10	6,82	11,53
GR+6,3	15,07	3,28	4,65	13,77	15,10	17,05	23,64
GR-6,3	84,93	3,28	76,36	82,95	84,91	86,23	95,35
GR+1	41,27	6,92	20,45	38,54	40,97	43,75	59,96
GR-0,15	35,37	7,76	15,30	32,10	35,74	40,02	54,67
GR-0,106	28,20	7,67	9,09	25,41	28,09	33,82	44,18
TML	12,47	1,73	9,31	10,74	13,43	13,88	14,74

Fonte: o autor.

Aqui, retoma-se o raciocínio da seção 2.3.4 em relação à cobertura dos intervalos entre os valores extremos assumidos pela variável resposta. Também na Tabela 7, é possível notar que os valores mínimo e máximo para a variável resposta no conjunto de dados teste são 9,31 e 14,74.

Desta forma, ainda sobre a variável resposta ($Y = TML$), conclui-se que os valores extremos dos 112 dados iniciais (9,05 e 14,79), sobre os quais foi estimado o modelo de regressão da equação 17, contém o intervalo dos valores extremos das 72 linhas de dados do conjunto teste (9,31 e 14,74), representado na Tabela 7. Este fato será importante na discussão de validação do modelo proposto em seções posteriores.

3.2. Análise de eficácia e validação do modelo preditivo

Para validação do modelo proposto e avaliação da capacidade preditiva, foram utilizados dados de 72 embarques realizados ao longo do primeiro semestre do ano de 2021, reunidos no já citado conjunto de dados teste e que foram apresentados na Tabela 7. Aplica-se a eles, portanto, o modelo de regressão linear múltipla proposto na seção 2.3.2, para avaliação dos valores preditos para a variável resposta.

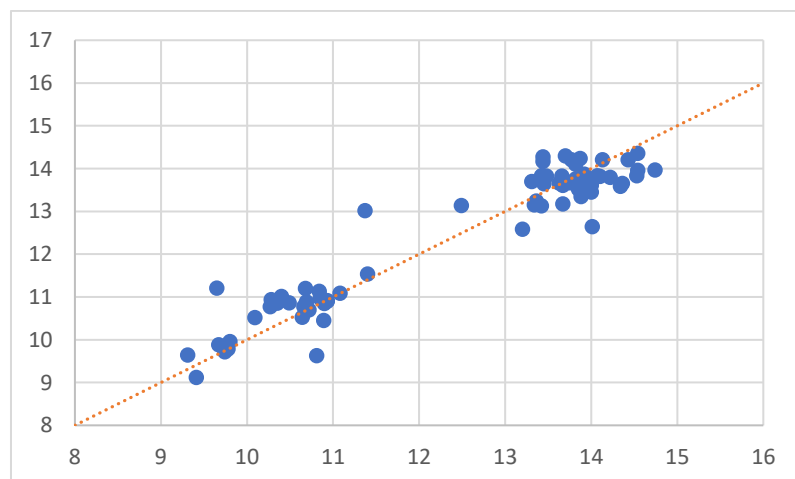
As estatísticas descritivas que comparam os valores reais, obtidos em laboratório, com os valores preditos pelo modelo estão dispostos na Tabela 8. Numa primeira observação, os números obtidos pelo modelo se mostram aderentes aos resultados empíricos.

Tabela 8: Estatísticas descritivas para variável resposta – TML real x predito

Variável	Média	Desvio Padrão	Mínimo	1º quartil	Mediana	3º quartil	Máximo
TML real	12,47	1,73	9,31	10,74	13,43	13,88	14,74
TML predito	12,49	1,60	9,12	10,91	13,30	13,83	14,36

Fonte: o autor.

O gráfico comparativo dos valores da variável resposta ($Y = \text{TML}$) obtidos pelo modelo *versus* os resultados de laboratório se encontram na Figura 6, bem como um gráfico de dispersão entre valores reais (abscissas) e estimados pelo modelo (ordenadas).



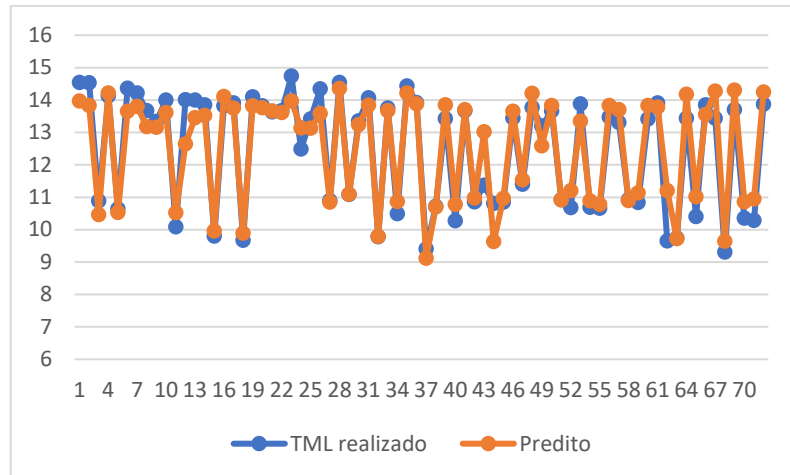


Figura 6: Gráficos de dispersão e de linhas do TML – Predito x Reais

Como explicitado para um dos gráficos da seção 2.3.4, observam-se dois grupos no primeiro gráfico da Figura 6. Tal fato se deve, novamente, à existência de dois tipos de produto de finos de minério de ferro na companhia, que possuem diferentes patamares de umidade e TML.

O modelo testado apresentou erro médio (ME) de 0,03, com viés maior para os números obtidos pelo modelo, quando comparados aos números do laboratório. Para a medida erro médio absoluto (MAE), o resultado foi 0,38.

Pelos gráficos da Figura 6, observa-se que o modelo foi capaz de descrever a tendência de comportamento do TML – com boa eficácia – para o intervalo estudado, tendo bom desempenho, inclusive, para valores mais próximos aos extremos da série de dados.

No gráfico de dispersão, nota-se que uma relevante quantidade dos pontos ficou próxima à linha de referência, como seria idealmente esperado do modelo. Paralelamente, o gráfico de linhas corrobora a eficiência do modelo em replicar a tendência de comportamento dos valores empíricos quando comparados às predições.

Outro ponto interessante reside no intervalo de cobertura entre os extremos da variável resposta. Como citado anteriormente, os valores de Y do conjunto de dados teste tem seus valores contidos no intervalo do conjunto de dados de concepção do modelo. Desta forma, toda a cautela necessária ao fazer extrapolações deixa de ser determinante na utilização do modelo, pois sua equação foi determinada num intervalo

amplo e representativo da realidade operacional. A potencial – e preocupante – influência da multicolinearidade na estimação dos parâmetros β fica, portanto, num segundo plano para o modelo, gerando maior confiabilidade na previsão do TML nos intervalos aqui expostos.

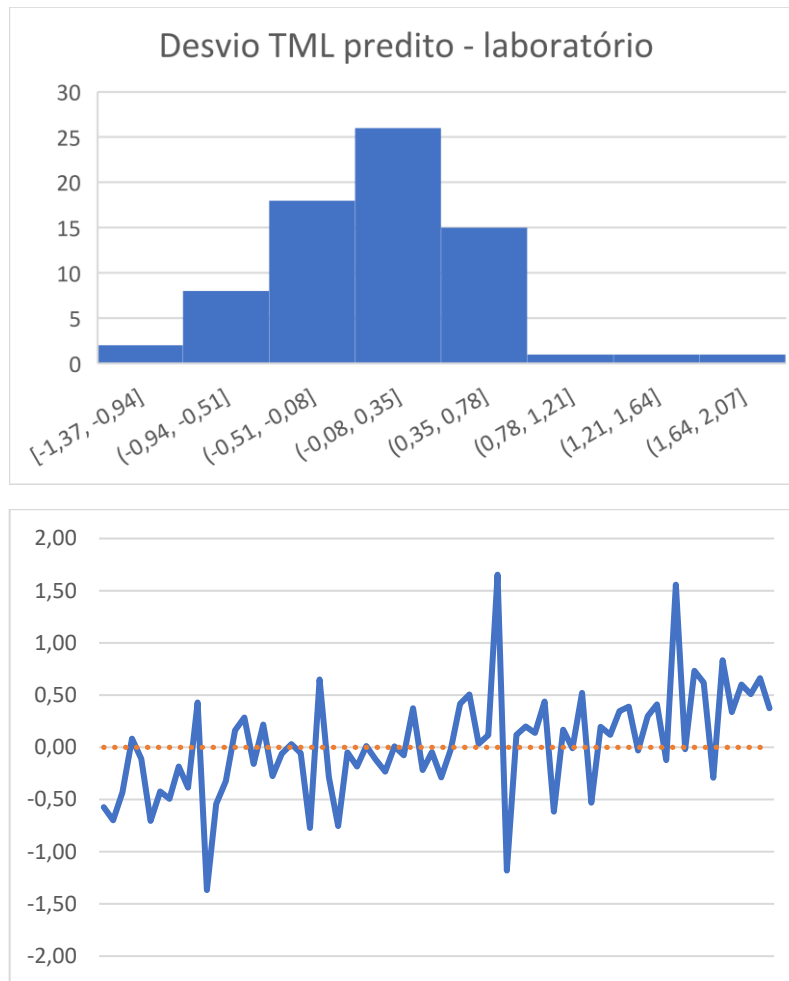


Figura 7: Histograma e gráfico de linha dos desvios TML Predito – Laboratório

Complementarmente, na Figura 7, observa-se o histograma e gráfico de linha para os desvios entre o TML predito em relação ao encontrado no laboratório, obtido por meio do método empírico de referência.

Conforme abordado na introdução deste texto, o ensaio de compactação que gera o resultado para o TML tem repetibilidade validada em 0,2 p.p., quando avaliada a diferença entre dois testes.

O modelo proposto produz $ME = 0,03$ e $MAE = 0,38$. Dos 72 dados testados pelo modelo, 45 deles foram preditos com erro, em módulo, maior de 0,2 p.p., algo em torno de 63%.

Porém, como o próprio método pode inserir um viés de 0,2 p.p., entende-se que, na prática, dada a simplicidade do modelo e de sua potencial aplicação, um erro acumulado em 0,6 p.p. (viés do método somado ao MAE, com arredondamentos) seria viável do ponto de vista operacional, por já ser útil no sentido de fazer uma avaliação preliminar da carga. Neste novo nível de crítica, apenas 15 dados ficaram fora da faixa de validação, ou seja, o modelo conseguiria produzir um resultado assertivo em quase 80% das simulações. Um resumo deste raciocínio está descrito na Figura 8.

Dif. Predito - Lab (p.p.)	Abaixo da faixa	Aderência	Acima da faixa	Dados Totais
Dif < -0,2 ou Dif > 0,2	21	27	24	72
	29,2%	37,5%	33,3%	100%
Dif < -0,6 ou Dif > 0,6	7	57	8	72
	9,7%	79,2%	11,1%	100%

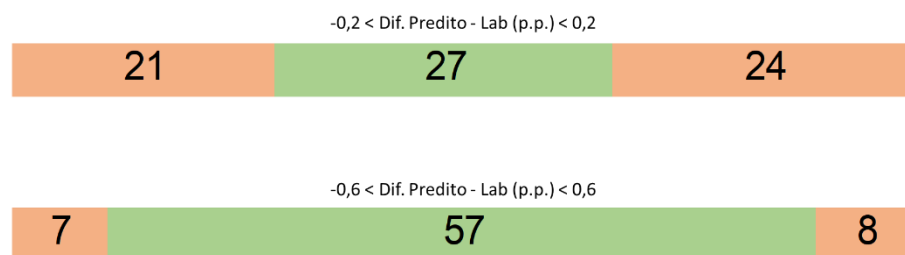


Figura 8: Avaliação prática do modelo – critérios de validação

Os 20,8% de resultados não assertivos do modelo (ver Figura 8) seriam, de fato, falhas no propósito de guiar o processo decisório das operações. Por outro lado, considerando a relativa simplicidade da ideia e de sua execução a partir de dados disponíveis na rotina, sem necessidade de alocação de qualquer recurso adicional, entende-se que este é um risco que vale a pena assumir, principalmente comparando a um cenário anterior totalmente empírico, via ensaios laboratoriais, sem qualquer ferramenta estatística a serviço dos analistas.

Além disso, é possível trazer em estudo futuros novas adequações à ideia, seja por meio da disponibilização de novos dados e parâmetros para enriquecimento do modelo ou mesmo pela aplicação de novos métodos, como ferramentas mais robustas

do ponto de vista computacional, que exigem mais recursos e acompanhamento, mas que podem gerar resultados ainda melhores.

4. CONCLUSÕES

Medir e controlar TML e umidade, na prática, é uma tarefa dominada pela indústria de mineração. A previsão assertiva do TML para cargas pontuais, por outro lado, carece de aplicação de estudos teóricos e empíricos mais aprofundados e uso de recursos estatísticos, além de, eventualmente, aportes financeiros em tecnologia.

Dado o objetivo deste trabalho de obter um modelo de predição do TML a partir de números já disponíveis na rotina, entende-se que a discussão aqui exposta corrobora a tese de que os parâmetros físico-químicos analisados nos laboratórios são, potencialmente, uma promissora fonte para desenvolvimento de preditores em diversos métodos.

Aqui, especificamente, obteve-se um modelo de regressão linear múltipla que, de forma simples, facilmente replicável e de baixo custo, prevê um valor para o TML de uma carga potencialmente embarcável em navios graneleiros.

Caso a matriz mineral ou produtos de finos de minério de ferro da indústria sofram alterações significativas em suas composições, o modelo pode ser novamente proposto a partir de novos dados e ser retreinado quantas vezes forem necessárias.

Outro ponto interessante é que, dada sua arquitetura simplificada e a grande difusão do *software* Minitab no mercado, não se torna um problema o uso da ferramenta no ambiente digital das grandes corporações, que costumam dispor de políticas de segurança de informação que podem ser bastante rígidas. Ainda que fosse escolhido outro *software*, pago ou gratuito, bastaria estar em acordo com as políticas internas da empresa para garantir a viabilidade da solução.

Em suma, compreende-se que modelos estatísticos clássicos, de maior ou menor complexidade, bem como novas ferramentas computacionais de mercado ou de desenvolvimento interno das companhias, podem ser a chave para alavancar o desempenho de diferentes indústrias. A oportunidade reside no uso de dados que, por vezes, existem em enormes quantidades, mas acabam subutilizados na rotina e deixam de contribuir na tomada das melhores decisões dentro das empresas.

REFERÊNCIAS

BURNHAM, K. P.; ANDERSON, D. R. **Multimodel Inference**. Sociological Methods & Research, v. 33, n. 2, p. 261–304, 2004.

FERREIRA, R. F. **Modelos para previsão do limite de umidade para transporte marítimo de finos de minério de ferro - TML**. Tese (Mestrado em Engenharia Mineral) - Escola de Minas, Universidade Federal de Ouro Preto, Ouro Preto, 2019.

FERREIRA, R. F.; POLICARPO, D. L. V.; PADULA, V. P.; FERREIRA, M. T. S. **Limite de umidade transportável de minérios de ferro: aspectos regulatórios e técnicos**. Tecnologia em Metalurgia Materiais e Mineração, v. 14, n. 1, 2017.

IBRAM. **A importância do limite de umidade para transporte marítimo de minério**, 2021. Disponível em: <<https://ibram.org.br/noticia/a-importancia-do-limite-de-umidade-para-transporte-maritimo-de-minerio/>>. Acesso em 2 de abril, 2023.

IBRAM. **Mineração em números**, 2023. Disponível em: <<https://ibram.org.br/mineracao-em-numeros/>>. Acesso em 2 de abril, 2023.

IMO - INTERNATIONAL MARITIME ORGANIZATION. **IMSBC Code: International Maritime Solid Bulk Cargoes Code: Incorporating Amendment 05-19 and Supplement**. International Maritime Organization, 2020.

INTERCARGO. **Bulk Carrier Casualty Report**, 2021. Disponível em: <<https://www.intercarga.org/wp-content/uploads/2022/04/INTERCARGO-Bulk-Carrier-Casualty-Report-2021-1.pdf>>. Acesso em 2 Abril, 2023.

KUTNER, M. H.; NACHTSHEIM, C.; NETER, J. **Applied linear regression models**. Singapore: McGraw-Hill Education/Asia, 2018.

MINITAB. **Interpretar todas as estatísticas para Regressão dos melhores subconjuntos**. Disponível em: <<https://support.minitab.com/pt-br/minitab/20/help-and-how-to/statistical-modeling/regression/how-to/best-subsets-regression/interpret-the-results/all-statistics/#r-sq-pred>>. Acesso em: 6 maio. 2023.

MINITAB® Statistical Software. **Versão 21.4.0** (via aplicativo online), 2023. Acesso em <<https://app.minitab.com/>>

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística aplicada e probabilidade para engenheiros**. Rio De Janeiro. LTC Editora, 5 ed, 2012.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis**. New York, NY. John Wiley & Sons, 2015.