Foto **Divulgação**

Por *Gláucia Cuchierato*[1]

GDQM: GEODATA QUALITY MANAGEMENT

Parte 1 – Validação do acervo de dados históricos

(Série de artigos técnicos sobre a importância da qualidade da informação utilizada na declaração de recursos e reservas minerais, iniciada na edição ITM 103, sobre os componentes do GeoData Quality Management, metodologia de avaliação da qualidade de dados geológicos).

Na última edição da revista, discutiu-se como é iniciada a primeira etapa da metodologia GDQM (GeoData Quality Management), de **Validação do Acervo de Dados Históricos** – verificação da fonte dos dados, avaliação e gestão da materialidade, identificação e integração dos sistemas de gerenciamento.

1) ANÁLISE EXPLORATÓRIA DE DADOS

A análise exploratória, detalhada e definida por Cuchierato (2022) e desenvolvida por Castilho e Cuchierato (2022), tem como objetivo caracterizar os dados e a evolução das variáveis numéricas e categóricas, além de identificar os avanços e otimizações em técnicas e tecnologia que foram adotados ao longo do tempo.

Após a extração da base de dados do sistema oficial da empresa, todas as tabelas são analisadas, com verificação do preenchimento de linhas e colunas por tipo de dados: object (texto), float64 (número com ponto flutuante), int64 (número inteiro) ou datetime (data ou data/hora). Na sequência, são organizados os dicionários de dados que, a depender da quantidade de variáveis, pode ser apresentado na forma de lista completa, ou apenas de máximos e mínimos com a indicação de todas as possíveis escolhas de preenchimento para cada campo, como observado no Quadro 1.

Após a análise global dos dados, busca-se verificar as variações no volume e na taxa de aquisição de dados, geralmente representados pelo número de furos e metros executados, para dividir o conjunto de dados em períodos com características similares. As Figuras 1 e 2 ilustram um exemplo de apresentação dos furos e metros executados, por período.

Quadro 1: Exemplo de colunas, tipos de dados e dicionário da tabela "Collar"

Coluna	Quantidade de linhas		Tipo do dado	Valores únicos	Lista de escolhas / Variação
	Preenchidas	Nulas			
BHID	444	0	object	444	DDH_P05A a PP_F051X
PROJETO	444	0	object	3	EXP / A1 / BX
PROF_FINAL	444	0	float64	731	87,30 a 422,65
XCOLLAR	444	0	float64	622	0 a 3500
YCOLLAR	444	0	float64	586	0 a 1500
ZCOLLAR	444	0	float64	259	675,25 a 731,12
ANO	444	0	Int64	15	2001 a 2015

A divisão dos períodos é iniciada pelo entendimento da linha do tempo dos projetos e estudos realizados e consulta a profissionais da equipe técnica. Complementarmente, verifica-se se os períodos definidos coincidem com as principais diferenças de evolução de metodologia de aquisição dos dados para todos os parâmetros ao longo do tempo, tais como:

a) análise por titulometria, via úmida, XRF pastilha prensada, XRF pastilha fundida;

b) levantamento topográfico com GPS de baixa precisão, RTK de alta precisão; e

c) perfilagem – equipamentos com bússola e mecanismo de marcação de tempo, método magnético|não magnético, com sensores ópticos, single e multishot|giroscópio e acelerômetro, com buscador de norte.

Destaca-se que a evolução metodológica e tecnológica é inerente à curva de aprendizado temporal, e supõe-se que, sempre, foram adotadas e aplicadas as melhores práticas disponíveis em cada campanha exploratória e confirmatória.

Também é importante espacializar os dados, para uma melhor compreensão do avanço das pesquisas no tempo. A Figura 3 ilustra esse exercício, com um mapa 2D das bocas dos furos e descrição das fases de sondagem.

A análise exploratória pode ser feita com todos os 28 parâmetros de qualidade definidos por Batini et al. (2009). Para esta proposta, foram escolhidos alguns dos critérios mais usuais da avaliação dos dados, indicados no Quadro 2, para instruir a operacionalização da análise proposta.

Figura 1: Evolução da execução de furos

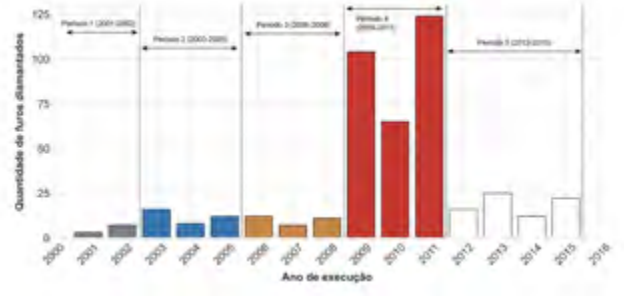


Figura 2: Evolução da execução de metros de sondagem

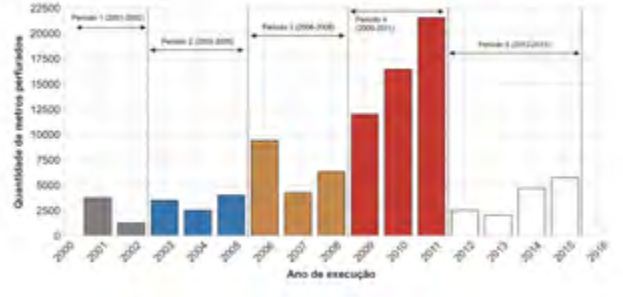
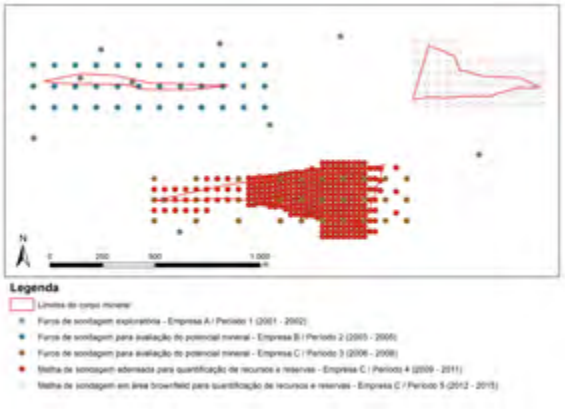


Figura 3: Espacialização dos furos executados por períodos



Quadro 2: Exemplos de parâmetros de análise exploratória recomendados

Critério	Definição
Singularidade	• A singularidade de um conjunto de dados é avaliada pelo menos por uma coluna (identificador único) ou conjunto de colunas (identificador composto). • Dados singulares são representados pela não repetição dos seus identificadores ao longo do conjunto avaliado.
Interpretabilidade e Clareza	• Busca avaliar a facilidade na leitura do conjunto de dados e a capacidade de transformação do dado em informação por parte do utilizador\usuário. • Este critério é composto por 4 partes: i)padronização das colunas de identificação; ii)presença de unidade da variável no nome da coluna; iii) presença de uma coluna equivalente numérico-categórica; iv)presença de comentários com vírgula.
Completude	• Razão entre o [número de valores nulos] e o [número total de dados armazenados] dentro de uma mesma tabela, coluna ou conjunto de dados. • Avalia o percentual de dados preenchidos em relação ao volume total de dados.

As premissas para avaliação da interpretabilidade e clareza das tabelas do banco de dados foram:

1) Colunas de identificação: Devem possuir o mesmo nome em todas as tabelas do banco de dados, como "BHID", "FROM" e "TO". Devem ser todas escritas em caixa alta, com exceção da primeira letra de elementos químicos e unidades de variáveis numéricas.

2) Unidade da variável: Quando se trata de dados numéricos é de extrema importância conhecer qual a unidade e método de análise (ppm, %, kg/ton, ICP, XRF), pois pode variar ao longo do tempo e deve ser preservada na estrutura do banco de dados. Idealmente, esta informação é a última dentro do nome de uma coluna.

3) Equivalente numérico-categórica: Em algumas situações, uma característica é classificada utilizando dados do tipo "object". Para a possibilidade de interpolação desta informação dentro dos softwares de modelagem, a melhor solução é existir uma coluna equivalente numérico-categórica, onde "0" é equivalente unicamente a "absent", ou dado ausente.

4) Comentários com vírgula: A coluna onde é permitida a inserção de texto descritivo não deve conter vírgulas, considerando que o arquivo exportado seja do tipo "comma separated values" (valores separados por vírgulas). A presença de vírgula nessas colunas irá impedir a leitura do arquivo de maneira correta ou, possivelmente, sobrescrever outros dados.

A análise da singularidade identifica, principalmente, dados e registros duplicados. Para tabelas onde cada furo ocupa apenas uma linha, como "Collar" e "Survey", a singularidade é avaliada apenas pela coluna de identificação do nome do furo ("BHID", por exemplo). Para as demais tabelas, nas quais existe mais de uma linha para cada furo, a singularidade é avaliada de modo a considerar as colunas "BHID", "FROM" e "TO", ou outras colunas que identifiquem e discretizem cada intervalo como único.

As análises de completude buscam evidenciar a disponibilidade de informação de modo quantitativo, conduzidas nos âmbitos:

i. completude em relação ao número de tabelas do banco de dados; e

ii. completude dentro de cada tabela, representada pela porcentagem de valores (linhas ou campos) nulos.

A completude do banco de dados avalia, ano a ano, a porcentagem de furos que foram descritos em cada tabela. Caso existam tabelas de preenchimento não obrigatório, pode-se subdividir o conjunto em "tabelas essenciais" (de preenchimento obrigatório) e "tabelas complementares". O número de tabelas e os tipos de dados que serão coletados devem ser planejados com base nas particularidades de cada área, idealmente antes do início das sondagens exploratórias, para definir o conjunto mínimo e completo de dados necessários a ser coletados de forma a obter bons resultados.

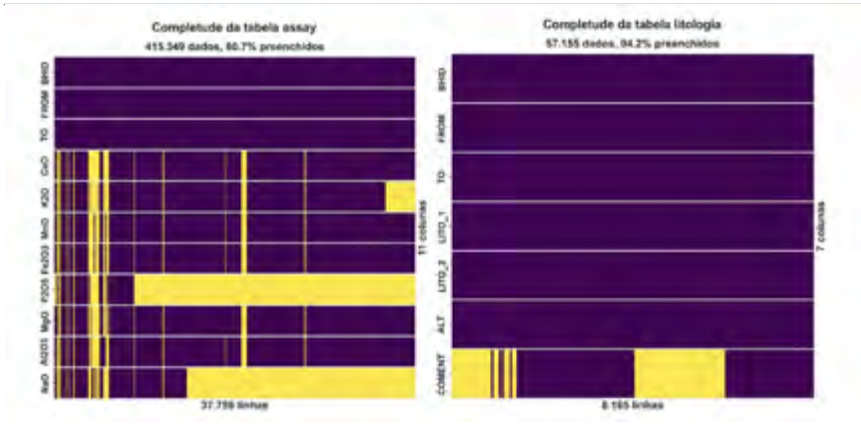
Dessa análise, verifica-se: 1) a tendência de aumento do preenchimento ao longo do tempo; 2) quais tabelas têm menores taxas de preenchimento, e os motivos; 3) a utilidade dos dados coletados; e 4) tabelas e colunas que podem ser descontinuadas para aumentar a eficiência da etapa de descrição.

Destaca-se que, dessa análise, não é possível atribuir qualidade ao dado, mas somente ao preenchimento e, em segundo plano, ao processo de descrição. A quantificação do preenchimento dos dados é representada no gráfico de completude (Figura 4) - em roxo com dados existentes e em amarelo com dados nulos.

Na próxima edição serão detalhadas as últimas etapas da validação do acervo – Aplicação dos testes de consistência, Atribuição de confiança e Identificação de áreas críticas.

Não percam o final desta etapa de análise!

Figura 4: Exemplos de diagramas de completude (Tabela "Assay" e "Lito")



*Veja a íntegra do artigo e referências bibliográficas em **inthemine.com.br***

1 Geóloga e Mestre em Recursos Minerais pelo IGc-USP, Doutora em Engenharia Mineral pelo PMI-EPUSP (Projeto: "O valor da qualidade da informação no processo de declaração de recursos minerais") e Diretora Executiva da GeoAnsata Projetos e Serviços em Geologia