DAYANA NIAZABETH DEL VALLE SILVA YANEZ

PREDIÇÃO QUANTITATIVA A PARTIR DE ANÁLISES POR *SELF-ORGANIZING MAPS*: ESTIMATIVAS DE SÍLICA REATIVA E ALUMINA APROVEITÁVEL EM BAUXITAS

Trabalho de Formatura em Engenharia de Minas do curso de graduação do Departamento de Engenharia de Minas e de Petróleo da Escola Politécnica da Universidade de São Paulo

Orientador: Prof. Dr. Cleyton de Carvalho Carneiro.

São Paulo

2015

DEDICATORIA

A Minha família, especialmente meus pais Ernesto e Thaely; minhas irmãs Stefany e Daniela; e meus avós Tania, Carlos, Ernesto e Yolanda; e ao Ricardo:

- Por ser a minha fonte de inspiração para continuar;
- Por me motivar e incentivar todos os dias com cada palavra, mesmo à distância;
- Por acreditar em mim e me fazer sentir que sou capaz de atingir todo desafio com sucessos;
- Por me impulsar a querer sempre mais desafios e dar o melhor de mim em cada um.

Eu dedico a minha primeira conclusão do trabalho a vocês, o qual marca o início de uma vida profissional.

Na esperança de dedicar lhes muitos mais sucessos ao longo deste maravilhoso caminho sem fim: a aprendizagem e o conhecimento.

AGRADECIMIENTOS

Agradeço a Deus por minha família; por colocar no meu caminho o meu parceiro durante anos; e os amigos. Todos estiveram presentes em cada um dos meus sucessos e fracassos durante esta longa e árdua jornada, mas certamente especial e inesquecível. Eles levaram-me a aproveitar todas as oportunidades, como foi o intercâmbio estudantil na USP, Brasil. Onde, além de fazer novas amizades, experiências, aprender uma nova língua, eu me levo uma porta aberta para novas oportunidades.

Da minha casa de estudos, a Universidade Simón Bolívar: Eu agradeço por meus colegas e por cada um dos professores com quem tive o privilégio de me formar, porque mais que académica, deram uma formação abrangente. Por me mostrar que, para ser um excelente profissional, antes devo ser um cidadão exemplar. Por me incutir o pensamento crítico para todos os desafios que a vida me apresenta e a superação de adversidades.

Aos Professores Marcio e Carina, que desde o início me deram a oportunidade e responsabilidade de realizar um trabalho de investigação e assim me contataram com um tutor excelente e dedicado durante o curso desta pesquisa: Prof. Cleyton. Obrigada por me guiar desta forma, foi especial e recompensador contar com seu profissionalismo, conhecimento e apoio incondicional.

Muito obrigada a todos por, de todas as formas possíveis,influenciarem para iniciar esta experiência, assumir a responsabilidade, me fazer sentir capaz de enfrentar o desafio e me motivar a fazê-lo cada vez melhor.

RESUMO

Análises geoquímicas são caracterizadas pela aquisição de medidas de múltiplas variáveis analíticas. Nesse sentido, a geração de amplos bancos de dados geoquímicos possibilita estudos a predição ou estimativa de valores analíticos ausentes ou complexos de medir. O beneficiamento da bauxita é uma etapa fundamental para a produção de alumínio, onde a determinação de teores de sílica reativa (SiR) e alumina aproveitável (AA) são fundamentais. Os métodos analíticos para obtenção desses teores apresentam limitações associados. à baixa repetitividade e reprodutibilidade dos resultados. A partir da predição quantitativa de valores provenientes da técnica não supervisionada *Self-Organizing Maps,* este estudo visa desenvolver sistematicamente a estimação de teores ausentes da composição química de amostras de bauxitas da base de dados de três projetos, a partir das variáveis: recuperação em massa (%) e teores (%) de AA; SiR; Al_2O_3 total; SiO_2 total; Fe_2O; TiO2; e/ou PF. Cada projeto foi submetido à exclusão parcial de valores de AA e SiR, em proporções de 20%,30%,40% e 50%, com a finalidade de investigar a técnica SOM como metodologia de quantificação de SiR e AA. Segundo os resultados obtidos na correlação e comparação dos valores preditos pelas análises SOM e os teores originais, foi possível avaliar a técnica SOM como ferramenta preditiva capaz de fornecer resultados analíticos satisfatórios com até 50% de exclusão de dados. Especificamente, os melhores resultados demonstram que a AA pode ser obtida por predição com maior correspondência que a SiR, tendo por base os parâmetros e variáveis envolvidas no estudo. A correspondência na natureza das amostras bem como a maior quantidade de variáveis analíticas inseridas também são quesitos que proporcionaram melhores resultados preditivos.

Palavras-chave: Predição Geoquímica Analítica; Self-Organizing Maps (SOM); Bauxita, Sílica Reativa (SiR), Alumina Aproveitável (AA).

ABSTRACT

Geochemical analysis provides the acquisition of multiple analytical variables measurement. Accordingly, the generation of large geochemical databases enables prediction studies or analytical estimate of missing values or complex measuring. The processing of bauxite is a key step in the production of aluminum, in which the determination of SiR and AA are very relevant. Analytical methods for achieving these concentrations have limitations associated with poor repeatability and reproducibility of results. Based on the quantitative prediction values from a unsupervised technique Self-Organizing Maps, this study aims to develop, systematically, the estimation of missing concentrations of the geochemical composition of bauxite samples of a database from three projects, from the variables: WT (%) and contents (%) of AA; SiR; total Al_2O_3; total SiO_2; Fe_2O; TiO_2; and / or PF. Each project was submitted to partial exclusion of AA and SiR values, in proportion of 20%, 30%, 40% and 50%, to investigate the SOM technique as quantification methodology of SiR and AA. According to the results obtained in the correlation and comparison of predicted values for SOM analysis and original values, it was possible to evaluate the use of SOM technique as a predictive tool capable of providing satisfactory analytical results with up to 50% of deleted data. Specifically, the best results demonstrate that AA can be obtained by prediction with the higher correspondence than SiR, based on the parameters and variables involved in the study. The match in the nature of samples and the greatest amount of embedded analytical variables are also parameters that provided better predictive results.

Key-words: Analytical Geochemical Prediction, Self-Organizing Maps (SOM), Bauxite, Reactive Silica (SiR), Available Alumina (AA).

LISTA DE FIGURAS

pág.

LISTA DE TABELAS

pág.

LISTA DE ABREVIATURAS

AA	Alumina aproveitável
Al_2O_3	Alumina
A_t	Número total de amostras
ASSOM	*Adaptative-Subspace* SOM
BMU	*Best Matching Units*
Corr	Correlação de dados

CP	*Components Plots*
$E_{20\%}$	Exclusão de 20% dos dados originais
$E_{30\%}$	Exclusão de 30% dos dados originais
$E_{40\%}$	Exclusão de 40% dos dados originais
$E_{50\%}$	Exclusão de 50% dos dados originais
ERP	Porcentagem de erro relativo
FASSOM	*Feedback Adaptive-Subspace SOM*
Fe_2O_3	Óxido de ferro
FSOM	*Feedback SOM*
IA	Inteligência Artificial
LCT	Laboratório de Caracterização Tecnológica
LVQ	*Learning Verctor Quantization*
L_1	Comprimento de treinamento robusto
L_2	Comprimento de treinamento calculado
Matriz-U	Matriz de distancia unificada
Max_t	Teor máximo
Min_t	Teor mínimo
M_{et}	Mediana de teor
M_t	Media de teor
n-D	n Dimensões
PCA	Padrão de Componentes Principais
PF	Perda ao Fogo
Q_e	Erro de quantificação final
R_{f1}	Raio final de dados robustos SOM
R_{f2}	Raio final de dados finos calculados SOM
R_{i1}	Raio inicial de dados robustos SOM
R_{i2}	Raio inicial de dados finos calculados SOM
RNA	Redes Neurais Artificiais
SiO_2	Óxido de silício
SiR	Sílica reativa
$Size_{som}$	Tamanho do Mapa SOM
SOM	*Self-Organizing Maps*
T_e	Erro topográfico final
TiO_2	Óxido de titânio
T_{LCT}	Teores originais obtidos no LCT
T_{SOM}	Teores obtidos por SOM
V	Número de Variáveis
WT	Recuperação em massa
XRF	Fluorescência de raios X

SUMÁRIO

pág.

1. INTRODUÇÃO

Atualmente, no campo da geofísica e geoquímica existem inúmeros avanços que permitem a aquisição de dados multivariados com alta densidade de amostragem. A análise desses dados, no entanto, exige maiores estudos metodológicos, sobretudo em relação à otimização e exploração das relações entre as diversas variáveis analisadas. Com a fase exploratória das bases de dados, torna-se possível inovar e aperfeiçoar na integração e interpretação, bem como prever e/ou estimar valores analíticos.

Fraser e Dickson (2007) afirmam que *Self-Organizing Maps* (SOM) pode ser considerado uma ferramenta de análise exploratória de dados, e o método pode ser utilizado para realizar grandes categorias de operações, tais como previsão ou estimativa, agrupamento, classificação, reconhecimento de padrões, e / ou redução de ruído[1].

Em relação ao exposto, torna-se possível implementar análises SOM como uma ferramenta alternativa auxiliar à predição de dados analíticos. Para abordar tais análises, serão exploradas medidas quantitativas relacionadas aos teores de Alumina Aproveitável AA (gibbsita) e Sílica Reativa SiR (caulinita) relativas a depósitos de bauxitas provenientes de diversas regiões do Brasil.

Inicialmente, foram feitas análises químicas em três projetos desenvolvidos no Laboratório de Caracterização Tecnológica (LCT) Departamento de Engenharia de Minas e de Petróleo (PMI) da Escola Politécnica da USP. Das variáveis obtidas nas análises químicas em amostras de bauxitas, os resultados de AA e Sir apresentaram baixa repetitividade e reprodutibilidade, além de altos custos analíticos, que tornam a realização de tais análises pouco viáveis [2]. Desse modo, técnicas de predição tais como SOM, possibilitam a estimativa de valores analíticos em dados multivariados.

A bauxita é um minério composto geralmente por gibsita, caulinite, e impurezas como o ferro, produzido pelo desgaste das rochas ígneas em condições geomorfológicas favoráveis. Por ser um agregado de varios minerais de aluminio, torna-se a matéria prima da qual é obtida a alumina e, consequentemente, o aluminio [3]. No cenário econômico do Brasil, o minério apresenta posição de destaque, detém a terceira maior reserva mundial e ocupa a segunda posição como país produtor [4].

O interesse econômico da bauxita depende de sua composição mineralógica, fundamentalmente das proporções de gibbsita e argilominerais contidos, sendo estes considerados deletérios para o aproveitamento econômico, que na prática é usualmente avaliada através de análises químicas específicas ou, mais raramente, de análises mineralógicas em estudos mais detalhados[5].

O conhecimento de um extenso banco de dados originado a partir de amostras de bauxitas, analisadas em três projetos, e medidas a partir da determinação de teores de AA e SiR, contêm elementos necessários e significativos para a predição de valores desconhecidos destas variáveis em outras amostras. A partir da técnica de analises de dados multivariados SOM, esta pesquisa visa gerar mapas auto-organizados, que representem as relações das amostras no complexo espaço n-D das variáveis. O mapa auto-organizado será a base para a predição das amostras sintéticas que visam se aproximar dos valores obtidos com as análises químicas de AA e SiR.

As análises SOM serão divididas em quatro fases para cada um dos projetos A, B e C. As amostras foram analisadas quimicamente a partir das variáveis: recuperação em massa (%) e teores (%) de (i) Al_2O_3 aproveitável; (ii) SiO_2 reativa;(iii) Al_2O_3total; (iv) SiO_2 total; (v) Fe_2O_3; (vi) TiO_2; e/ou (vii) Perda ao Fogo (PF). As fases constam da omissão parcial de valores analíticos em proporção de 20%, 30%, 40% e 50% de AA e SiR para as amostras. Os resultados preditos pela técnica SOM serão correlacionados com os valores químicos analíticos originais, e avaliados segundo estatística descritiva.

Uma vez demonstrada a correlação entre os dados originais e os preditos, as análises preditivas por SOM podem auxiliar aos usuários de análises químicas em bauxitas na obtenção de teores de AA e SiR com baixo custo material e pessoal. Os testes analíticos e probabilísticos nos referidos resultados trariam segurança e confiança para a utilização da ferramenta como fator de referência em outras análises.

Os experimentos desenvolvidos neste projeto visam abordar discussões relativas à validação de uma nova sistemática se obter teores de AA e SiR, com baixo custo operacional, sem comprometer os padrões de qualidade dos resultados. Além disso, os resultados visam promover o uso da técnica SOM como ferramenta preditiva capaz de fornecer resultados analíticos satisfatórios. Desse modo, os resultados que serão abordados a seguir abrem precedente a aplicação da técnica em estudos geoquímicos, geofísicos ou nas mais diversas

áreas, onde ocorram semelhantes incertezas ou necessidades em relação à predição, integração e interpretação de dados multivariados.

2. OBJETIVOS

2.1. Objetivo Geral

Desenvolver sistematicamente a predição de composição química/mineralógica em amostras de bauxitas, utilizando a técnica não-supervisionada SOM.

2.2. Objetivos Específicos

- Implementar e validar uma metodologia de quantificação de sílica reativa e alumina aproveitável em bauxitas brasileiras usando a técnica de análise espacial de dados SOM;
- Avaliar comparativamente as análises químicas convencionais e os valores de predição obtidos por meio da técnica SOM, confrontando-as em termos quantitativos bem como no que se refere à repetitividade em diferentes tipos de bauxitas;
- Direcionar parâmetros que possam auxiliar o uso da técnica de análises multivariado SOM para a predição, classificação, integração e interpretação de dados geofísicos.

3. REVISÃO BIBLIOGRÁFICA

3.1. Aspectos históricos da técnica SOM

Em 1984, o cientista Teuvo Kohonen desenvolveu uma monografia intitulada "Auto-Organização e Memória Associativa", que despertou o interesse de cientistas e pesquisadores da área de redes neurais. Tal publicação impulsionou os algoritmos de auto-organização, chamados de rede neural *Self-Organizing Maps* (SOM) e *Learning Verctor Quantization* (LVQ) a se tornarem mais populares. Os crescentes usos da técnica tornaram necessário estudo e abordagem mais criteriosa sobre a análise de padrões estatísticos, bem como maiores detalhamentos sobre a técnica SOM. Devido a isso, em 1998 Kohonen realizou uma segunda investigação, que teve início em 1981, com foco principal em SOM, e que deu origem à publicação do artigo científico The Self-Organizing Maps (1998), em seguida, três edições do livro com o mesmo nome [6].

A partir da concepção da base conceitual, as análises SOM foram utilizadas em diversas e extensas pesquisas ao longo da última década. Alguns destes estudos serão abordados abaixo, sobretudo aqueles com enfoque na aplicação das análises SOM em pesquisas geocientíficas, envolvendo geofísica e geoquímica.

Strecker e Uden (2002) usaram o princípio não supervisionado da análise SOM, de que a rede neural é livre para procurar, reconhecer e classificar padrões estruturais em um campo vetorial n-dimensional que abrange todos os conjuntos de dados 3D de atributos sísmicos em 3D [7]. Os autores concluíram que os mapas gerados por SOM proporcionam uma oportunidade para interpretações geológicas de grandes volumes de dados sísmicos 3D, quando a estratigrafia é largamente mascarada nos dados empilhados. Além disso, os autores conseguiram a distinção de características físicas em escala de reservatório, além de demonstrar a importância da heterogeneidade de reservatórios e otimização subsequente da respectiva exploração através de furos horizontais de poços. Tais resultados proporcionaram relevantes contribuições geofísicas [8].

Penn (2005) abordou a necessidade de visualização adequada para os conjuntos de dados de alta dimensão. A pesquisa foi exemplificada com os principais elementos geoquímicos, bem como com dados hiperespectrais [9].

Freser e Dickson (2007) propuseram abordagem computacional baseado em SOM para compreender e sintetizar a quantidade de dados exploracionistas adquiridos em estudos geológico, geoquímico e geofísico. Eles introduzem a análises SOM como capaz de relacionar e ajudar no processo de criação de conhecimento a partir de dados complexos e díspares com a finalidade de integrar e interpretar os dados. Além disso, os autores enfatizaram que SOM é uma ferramenta de análise exploratória não supervisionada e orientada a dados, a partir da qual os padrões resultantes, fronteiras e relações são internamente derivados [1].

Seguidamente, Carneiro *et al.* (2012) demonstraram a eficácia do uso SOM como uma ferramenta para análise de dados geofísicos tendo em vista a geração de mapas geológicos semiautomáticos.Para tanto foram desenvolvidas análises de classificação de variáveis geofísicas adquiridas a partir de levantamento aéreo realizado sobre a Amazônia brasileira. As análises SOM permitiram identificar e mapear de maneira confiável informações geofísicas relacionadas com as diversas unidades geológicas. As análises foram realizadas a partir dos

dados magnetométricos e gamaespectrométricos e foram relacionados a processos geológicos presentes na área[10].

Cracknell *et al.* (2015) usaram SOM para descrever as características geofísicas e mineralógicas do regolito e rocha. A aplicação de aprendizagem estatística em diversas camadas de dados de sensoriamento remoto terrestre, em escala continental, permitiu-lhes explorar as múltiplas influências da rocha-mãe, do clima, biota, paisagem e tempo no desenvolvimento do regolito e suas propriedades. Os autores apresentaram um exemplo de modelagem geocientífica interdisciplinar, realizada a partir de dados geofísicos e geoquímicos[11].

3.2. Modelo Neuronal

O sistema nervoso é organizado em termos de um número imenso de unidades elementares chamadas neurônios, dispostos em constelações funcionais ou conjuntos de acordo com os contatos sinápticos que fazem umas com as outras [12]. Além disso, os neurônios são estruturas de ocorrência natural que vieram em uma imensa variedade de tamanhos e formas independentemente de seu funcionamento fisiológico e psicológico.

Entretanto, a neurofisiologia apresenta a célula neural como um sistema dinâmico complexo controlado pelos sinais neurais, campos elétricos pequenos e numerosos transmissão químicas e moléculas transmissoras de mensagens [6].

3.2.1. Redes Neuronais Artificiais

Hoje em dia, o principio básico dos neurônios no cérebro humano, é usado em múltiplas aplicações de controle, baseado nas redes neurais artificiais (RNA's), com a finalidade de criar inteligência autônoma. Essas redes, a saber, teriam a capacidade de aprender, fazendo o possível para alcançar um elevado grau de autonomia e com a capacidade de aproximar funções altamente não-lineares, o que permite a construção de sistemas complexos de modelagem geral.

A fabricação de modelos, universalmente, busca o desenvolvimento de uma imagem sintética, consistente e instrutiva da natureza [13]. Estes modelos, por sua vez, consistem em um *set* finito de variáveis e múltiplas interações quantitativas, dispostos a descrever, mediante processos, estados e sinais em um sistema real com a finalidade de analisar, descrever, explicar, prever e simular.

A Inteligência Artificial (IA), por definição, é a capacidade que as máquinas têm em realizar atividades que exijam inteligência humana. Kohonen [6] descreveu três categorias principais que dominam a pesquisa de redes neurais artificiais: (i) os modelos de transferência de estado;(ii) os modelos de transferência de sinal; e (iii) a aprendizagem competitiva.

> Por uma questão de fato, os modelos de transferência de estado são casos especiais de circuitos de relações não-lineares, e os modelos de transferência de sinal são muito semelhantes às expressões em teoria da aproximação matemática. (...) A aprendizagem competitiva é relacionada à quantificação vetorial. Em todos os campos tradicionais existe abundância de resultados matemáticos que poderiam ser transferidos para as investigações de redes neurais [6].

É importante mencionar que a IA não é restrita só aos campos de modelagem e controle.Os temas que envolvem suas teorias e princípios têm atraído pesquisadores desde 1956 e, na atualidade, são utilizados nas mais variadas e diversas aplicações. Um dos exemplos relacionados envolve a estimativa de parâmetros não mensuráveis como uma alternativa para os observadores convencionais e sensores de hardware em sistemas de processos químicos. A afinidade ocorre dada a formulação simples, capacidades de adaptação e de requisitos mínimos de modelagem inerentes à IA [14].

Baseado no princípio de que a rede neural é livre para procurar, reconhecer e classificar padrões estruturais em um campo vetorial n-dimensional, IA é usada nos estudos não supervisionado da análise SOM[8]. Em geral, as RNA's demonstram-se como uma solução emergente para reconhecimento de padrões.

3.3. Self-Organizing Maps

> O *Self-Organizing Maps* (SOM) é uma técnica eficaz para a visualização dados de elevada dimensionalidade. Os princípios do SOM envolvem o mapeamento ordenado de uma distribuição de alta dimensão em uma rede regular de baixa dimensão. Deste modo, SOM é capaz de converter relações estatísticas complexas, não-lineares entre os itens de dados de alta dimensões em relações geométricas simples apresentadas via *display* de baixa dimensão. Como ele comprime informações, preservando as relações topológicas e métricas mais importantes dos itens de dados primários no visor, a técnica pode também ser utilizada para produzir alguns tipos de abstrações. Estes dois aspectos, visualização e abstração, podem ser utilizados de diversas formas nas tarefas complexas, tais como

> análise de processo, percepção automatizada, controle e comunicação [15].

Baseado em um modelo neural, na quantização de vetores e medidas de similaridades, os algoritmos usados por SOM analisam e integram dados em n-dimensões, cada uma representando uma variável de entrada, contínua ou categórica, com a finalidade de gerar um mapa em duas dimensões que permita interpretar dados complexos e díspares. Por esta razão, é fato que SOM tem uma abordagem na análise, integração, visualização e interpretação de dados.

SOM pode ser usado para predição, estimação, agrupação de padrão de reconhecimento e redução de ruído. Porém, para representar os múltiplos valores das amostras de entrada em um espaço 2D, torna-se necessário um treinamento dos vetores-nós, a partir de medidas de similaridade vetorial, seguindo regras matemáticas como produto escalar, cosseno, distância euclidiana, dentre outras.

3.3.1. O Subespaço Adaptativo SOM (ASSOM)

O principio do subespaço adaptativo de SOM, do inglês *Adaptative-Subspace* SOM (ASSOM) foi introduzido por Kohonen em 1995 como um tipo especial de SOM, com a finalidade de resolver uma das limitações de estes analises: as unidades dos mapas obtidas são sensitivas para alguma classe de padrões elementares que ainda não podem ser consideradas como características invariáveis.

A saber, ao definir filtros que correspondem aos vetores, se gera um padrão de subespaços, e podem ser excluídos alguns grupos pela transformação automática do vetor. Por essa razão, no ASSOM, as diferentes unidades de mapas se desenvolvem em filtros de muitas características básicas não variáveis, onde a unidade do mapa não está descrita pela longitude de um único vetor, mas esta é destinada a representar um subespaço linear compreendido pelo vetor básico adaptativo, processo alcançado pelo tipo de aprendizagem, que deve modificar todos os vetores base que definem um subespaço definido.

Assim, o ASSOM, baseado da combinação do SOM e o método de subespaço, é uma alternativa ao método de análise de padrão de componentes principais (PCA) de geração de recurso, verificado em diferentes estudos e abordagem neural para PCA [16]. Além disso, ASSOM pode gerar filtros de recursos espacialmente ordenados, devido às interações entre os

módulos de processamento [17]. É importante mencionar que as equações matemáticas dos filtros não necessitam ser fornecidas a priori; os filtros e suas variações se moldam automaticamente em resposta à transformação típica que ocorre na observação [6].

Efetivamente, ASSOM não reproduz padrões particulares, mas sim a transformação que ocorre nos seus próprios núcleos. Portanto, funciona a partir de um algoritmo que difere dos demais algoritmos de rede neural, com a finalidade de reproduzir um número de múltiplas características invariantes.

A entrada para uma matriz ASSOM é tipicamente uma sequência de padrão vetorial treinado para abranger certo subespaço, normalmente de elevadas dimensão. Tal entrada será adaptada para capturar a transformação nele codificada [18]. A função mais essencial de ASSOM, portanto, é o aprendizado competitivo de episódios, o que não denota um padrão, mas sim uma sequência de padrões construída a partir das unidades do mapa. Isto denota um processo que só ocorre neste modelo de rede neural.

3.3.2. *Feedback* do Espaço Adaptativo SOM

O *Feedback* SOM (FSOM) é um uma variação do modelo SOM, mas com o mesmo tipo de aprendizagem de vetores. Desse modo, torna-se possível a classificação temporal, a partir da expansão ou contração dos padrões de espaço temporal, segundo foi demonstrado por [19]. Estes autores apresentam a estrutura de FSOM como clara e simples, onde a informação de mapeamentos realizados é reincorporada ao espaço de entrada (*input*) com a finalidade de processar a informação temporal.

O ASSOM, portanto, pode ser considerado um método capaz de classificar os padrões adaptados espacialmente. Por outro lado, o FSOM, apesar das vantagens relacionadas ao processamento dos dados, possui uma complexa classificação dos respectivos padrões. Assim, torna-se necessário introduzir o termo Subespaço Adaptativo do *Feedback* SOM, do inglês *Feedback Adaptive-Subspace* SOM (FASSOM), o qual consiste na combinação dos dois modelos. Esta combinação é produzida a partir das funções de padrões de reconhecimento de saída e entrada dirigida, que podem ser feitas de forma adaptativa.

Por conseguinte, a arquitetura de reconhecimento de padrões definida como um FASSOM é uma variação do modelo ASSOM, proposto com a finalidade de gerar resultados da classificação obtidos por algum algoritmo para o padrão de reconhecimento. Para esta

variação, ASSOM só proporciona as características de entrada e o processo de aprendizagem adaptativo de ASSOM nos subespaços é controlado por informação de maiores niveles [6].

A ideia principal do FASSOM é que o fator de uma taxa de aprendizagem do algoritmo não supervisionado possa ser escalonado. Assim, altos valores serão obtidos se a classificação estiver correta e baixos valores serão obtidos se a classificação estiver errada. Desse modo, as ações do FASSOM são regidas pela estratégia de *feedback*. Naturalmente, o sinal resultante pode também se tornar negativo quando a classificação de treinamento está errada. Esta seria, portanto, uma punição tal como ocorre no aprendizado supervisionado em geral[6].

3.4. Aprendizagem da Quantização vetorial

A aprendizagem dos vetores é baseada em um treinamento de vetores em nó (*node vectors*) que pode ser descrito como um processo de dois passos iterativos: (i) competitivo; e (ii) comparativo. Esses processos são aplicados para cada amostra de entrada até que os vetores sejam capazes de representar a estrutura e os padrões de as amostras de entrada a partir das respectivas similaridades [1].

O passo competitivo, tendo como base medidas de similaridades vetoriais, é feito pela comparação da amostra de entrada e todos os vetores dentro de um raio particular. O vetor mais semelhante, ou vencedor *(winning)* terá suas propriedades modificadas de forma percentual, sendo que suas características buscarão semelhança junto à amostra de entrada. Já no passo cooperativo, todos os vectores dentro de um determinado raio do vetor vencedor são também modificados, de modo a que as suas propriedades também são alteradas por uma percentagem para procurar assemelhar-se à amostra de entrada em questão.

Ao final, com a aplicação desses passos a cada dado de entrada repetidamente, os vetores iniciais, agora vetores em nó, irão representar a estrutura original dos dados de entrada de forma automática e organizada. Sem precisar parametrização ou supervisão, os *seed-vetors*, conhecidos como "*Best Matching Units*" (BMUs), geram um mapa auto-organizado em duas dimensões, feito a partir de dados multivariados complexos.

Os BMUs são projetadas para o hipersuperficie envolvente e transformadas para produzir a representação dos dados no mapa auto-organizado. Uma vez calculada, o mapa pode ser visualizado de muitas maneiras [20]. Algumas destas são *components plots (*CP*)*;

um *K-means* do vetor dos valores de BMU;e a "matriz de distância unificada" (Matriz-U), utilizada neste projeto.

3.4.1. Visualização e interpretaçao do mapa: Matriz de distancia unificada (Matriz-U)

A visualização Matriz-U mostra a similaridade relativa, em termos de distância Euclidiana, entre vectores BMU adjacentes, representadas como nós no mapa auto-organizado. Assim, cada BMU na matriz contém a medida de similaridade do vector segundo as características entre as unidades de neurônios adjacentes [21].

A matriz-U usa vetores de código SOM para gerar uma visualização 2D de dados multivariados, conseguido pelas relações de propriedades topológicas entre os neurônios após a processo de aprendizagem. Este algoritmo gera uma matriz, na qual cada componente tem uma distância entre dois neurônios adjacentes. Altos valores da matriz representam uma região de fronteira entre os neurônios, e baixos valores representam um elevado grau de semelhança entre os neurônios.

O método Matriz-U tem a vantagem de apresentar de maneira mais clara as estruturas complexas não-lineares. Em uma matriz-U são descritas tanto as distâncias dentro de um aglomerado quanto a forma das distâncias entre os diferentes neurônios [20].

Dentro da representação demonstrada pela Matriz-U, o tamanho dos hexagonosvariam de acordo com o número de amostras de entrada, representada pelo respectivo BMU. As variações são mostradas em uma escala de cores da temperatura.Desse modo, o azul (mais frio) indica similaridade e proximidade entre nós adjacentes,enquanto as dissimilardade, maior distância são representadas com amarelos, laranjas e vermelhos (mais quente) [10].

Depois do treinamento e aprendizagem dos vetores em nó, e o agrupamento do SOM segundo o apresentado na visualização Matriz-U, é necessário conhecer as bases para uma ótima interpretação dos mapas. Entretendo, os vetores treinados de SOM são eficientemente utilizados para a visualização, além de que é fundamental no processo de agrupamento para reduzir a complexidade computacional [22].

O gráfico de vetor quantizado ordenado, o qual é uma superfície elástica gerada pelas projeções não-lineares e os métodos de agrupamento da análise SOM,devem ser interpretados

com base na estatística, seguindo a distribuição dos dados de entrada, onde cada nó tem duas direções principais encontradas pelas diferenças entre os vetores nas proximidades.

Em primer lugar, cada nó ou BMU apresentado no mapa é um agrupamento inicial das amostras de entrada, e no mapa Matriz-U2D são percebidas as mudanças de densidade, produzidas pelas estruturas locais na topologia, preservando sempre a projeção do espaço dos dados de alta dimensão.

A contribuição das variáveis é significativa na estrutura agrupada se os componentes da uma área local de SOM são grandes, segundo as diferenças vetoriais. Isso proporcionaria um amplo poder explicativo no domínio dos valores. Além, o poder discriminatório das variáveis é mais evidente nos extremos dos grupos, onde se observam variáveis com menor semelhança em relação às amostras próximas.

É importante mencionar que este tipo de interpretação visual é aplicável para alta quantidade de dados, a saber, matrizes extensas. Ao contrario, a resposta Matriz-U seria complexa e difícil de interpretar, além de fomentar um maior erro associado.

3.5. Aplicação das Análises SOM como Estimador

As análises SOM são catalogadas como associativas e estimadoras, devido ao fato de que todos os padrões dos mapas de saída (mapas auto-organizados) estão relacionados com os domínios de cada dado de entrada. Desse modo, o mapa gerado pode ser classificado em simétrico ou assimétrico.

No mapeado simétrico (Associativo), cada componente do mapa SOM (saída) corresponde a um sinal de entrada não condicionado, com as vantagens de rápida aprendizagem para ordenar e uma distribuição de tamanho uniforme. Porém, a função simétrica pode gerar defeitos topológicos globais no centro do mapa, devido ao fato de que os dados de saída replicam os defeitos dos dados de entrada. Assim, a função precisaria de muitas iterações para corrigir o mapa ruidoso.

Em comparação com a função simétrica, a vizinhança assimétrica acelera o processo de correção, mesmo na presença do defeito. No entanto, esta assimetria tende a gerar um mapa distorcido. Isto pode ser suprimido pela função de vizinhança assimétrica [23].

3.6. Estudos de Sílica Reativa (SiR) e Alumina Aproveitável (AA) nas Rochas de Bauxita

A bauxita, normalmente apresentada como uma mistura de gibbsita Al $(OH)_3$ e caulinita $Al_2Si_2O_5$ $(OH)_4$, é a matéria prima a partir da qual é obtida a alumina. Através da transformação da alumina é obtido alumínio (Al)[24]. Bauxitas de alto teor geralmente apresentam de 40 a 50% de Al_2O_3 e de 10 a 24% de Fe_2O_3.

Por essa razão, a bauxita, formada de rochas aluminosas que mobilizam minerais, elementos e substâncias químicas, é considerada a principal matéria prima utilizada na indústria de alumínio. O metal é um excelente condutor de calor e eletricidade, tendo a leveza como sua maior vantagem.

A composição mineralógica da bauxita, considerada como simples, é estimada e avaliada através de análises químicas de teores totais de Al_2O_3, SiO_2, Fe_2O_3 e TiO_2, usualmente por fluorescência e difração de raios X, e teores específicos de Al_2O_3 aproveitável e SiO_2 reativa, por via úmida. As estimativas mineralógicas são feitas nas bauxitas baseados nas seguintes suposições [5]:

- Teores de AA se relacionam a teores de gibbsita;
- Teores de SiR se relacionam a teores de argilominerais, em especial caulinita;
- Os teores de óxi-hidróxidos de ferro assumem os valores dos teores de Fe_2O_3;
- Os teores de óxidos de titânio assumem os valores obtidos para TiO_2;
- O teor de quartzo está relacionado à sílica não reativa (SiO_2 total - SiO_2 reativa).

Apesar do entendimento sobre a simplicidade da composição química das bauxitas, o processo para a obtenção de dados analíticos não obedece ao mesmo entendimento. As técnicas e ferramentas utilizadas para a caracterização química das bauxitas apresentam limitações que causam baixa repetitividade e reprodutibilidade, aliados aos altos custos analíticos.

A fluorescência de raios X (XRF) é uma das técnicas mais eficiente para caracterização química de materiais. A XRF determina elementos de maneira qualitativa e quantitativa através dos comprimentos de onda e intensidades de emissões características [25].

No entanto, a técnica é pouco eficiente para analisar elementos de número atômico baixo, bem como e apresenta limitações na calibração. Desse modo torna-se necessária a comparação com padrões semelhantes às amostras e com teores já conhecidos.

Por outro lado, a difração de raios-X é um dos métodos mais utilizados na caracterização de materiais cristalinos. Através da técnica, torna-se possível o desenvolvimento de análises microestruturais, tanto no aspecto qualitativo quanto no quantitativo, para a obtenção das propriedades físicas da bauxita. No entanto, a quantidade de dano por espalhamento elástico útil é centenas de vezes maiores do que para os elétrons sem todos os comprimentos de onda e, portanto, as energias e exigências sobre o tamanho da amostra e número de partículas é muito maior [26].

É importante mencionar que a técnica de via úmida também apresenta limitações, tais como: erros analíticos, substituições de elementos dentro da estrutura cristalina de um mineral e presença de minerais com composições químicas similares. Isso torna o uso da técnica limitado, bem como acarreta em desvios na estimativa mineral [2].

4. MATERIAIS E MÉTODOS

4.1. Seleção de Amostras de Bauxita para as Análises SOM

As amostras de bauxitas selecionadas para as análises SOM são parte das bases de dados de três projetos compostos por particularidades litológicas diferentes. O confronto entre diferentes projetos teve por finalidade a obtenção da melhor representatividade e comparabilidade entre estes.

A estimativa e avaliação da composição mineralógica das amostras de bauxita foram desenvolvidas mediante análises químicas de teores totais de Al_2O_3, SiO_2, Fe_2O_3 e TiO_2, XRF, e teores específicos de Al_2O_3 aproveitável e SiO_2 reativa, por via úmida no Laboratório de Caracterização Tecnológica (LCT), do Departamento de Engenharia de Minas e de Petróleo (PMI) da Escola Politécnica da Universidade de São Paulo (EPUSP).

Efetivamente, foi priorizada a variabilidade química e composicional entre elevados e baixos teores de SiR e AA. Desse modo, a variabilidade da proveniência das amostras, e os distintos processos aos que foram submetidas, tais como ensaios de classificação e separação de minerais, viabilizaram uma extensa e diversa base de dados para realizar as análises SOM.

Os dados adquiridos do **Projeto A** contam com dez grupos compostos por sessenta e nove amostras cada um. As variáveis caracterizadas a partir de cada amostra são: os teores totais de Al_2O_3, SiO_2, Fe_2O_3 e suas porcentagens, e de recuperação em massa (WT).Com relação aos dados utilizados no **Projeto B** contam com os teores das variáveis: Al_2O_3, SiO_2, Fe_2O_3 e TiO_2 em valor, em porcentagem e Perda ao Fogo (%PF) para vinte e cinco grupos de amostras, subdivididos em seis grupos de oito amostras e o dezenove grupos de nove amostras. Por fim, o **Projeto C** conta com um grupo de setenta amostras, caracterizados em teores totais de: Al_2O_3, SiO_2, Fe_2O_3 e TiO_2, em porcentagem, e Perda ao Fogo (%PF)[2].

4.2. Procedimento Experimental

Uma vez selecionadas as amostras, foi necessária a preparação e organização destas com a finalidade de alimentar as análises de predição dos valores a partir da técnica SOM. Na preparação de amostragem foram excluídos valores aleatórios de AA e SiR para posterior estimativa pelas analises SOM. Finalmente, os valores preditos foram comparados com os valores de originais obtidos por análises químicas de teores por XRF ou via úmida no LCT.

4.2.1. Preparação das amostras

Com a finalidade de medir e avaliar a extensão das analises SOM, foram modificadas as tabelas de dados de amostragem com a omissão aleatória de valores de AA e SiR. Para cada projeto, foi filtrado aleatoriamente 20%, 30%, 40% e 50% do total de amostras, para a geração das novas tabelas de entrada a serem utilizadas na predição de valores. A tabela de dados, então, foi introduzida na plataforma SOM a partir do software SiroSOM®.

Nesse sentido, o espaço de dados foi introduzido de forma aleatória. Como variáveis de entrada foram utilizadas a ***recuperação em massa (%)***, ***teores (%)*** e ***recuperações metalúrgicas (%)*** das variáveis: (i) Al_2O_3 aproveitável; (ii) SiO_2 reativa; (iii) Fe_2O_3, (iv) TiO_2, e Perda ao Fogo (PF).

Tabela 1- Preparação de amostras para cada projeto

	At	E20%	E30%	E40%	E50%	V
Projeto A	690	140	210	280	350	4
Projeto B	219	48	72	96	120	7
Projeto C	70	14	21	28	35	6

4.2.2. Predição de valores a partir da Técnica SOM

A predição de valores de teores de AA e SiR das bauxitas foi desenvolvida segundo a adaptação de uma rutina proposta [10], onde os dados estimados pela técnica têm como base as distâncias entre os vetores disponíveis [1]. Para os dados com menor resolução espacial, o processo de estimativa tradicional é dado por substituição, onde os valores são produzidos a partir dos vetores das BMU's. Muitas vezes os conjuntos de dados resultarão em previsões tendenciosas, o que faz necessário a utilização de técnicas como do vizinho mais próximo [27].

A pesquisa contou, além da equipe envolvida, com o suporte técnico do instituto de pesquisa australiano CSIRO, com a colaboração do pesquisador Stephen Fraser (CSIRO), que presume na disponibilidade do software SiroSOM® para as análises em ambiente de SOM.

Uma grade hexagonal foi escolhida como formato de visualização; a superfície de um hiper volume toroidal foi utilizada para a projeção dos neurônios ou BMU's. Para a definição do tamanho de mapa auto-organizável resultante, foi utilizada a equação 1 (Eq. 1), onde [n] representa o número de amostras inseridas na plataforma SOM [28]. Dessa maneira, um tamanho de mapa foi escolhido como adequado para este estudo exploratório. Após a geração do mapa auto-organizado, foram produzidas as imagens da Matriz-U e CP.

$$\boldsymbol{Size_{SOM}} = 5x\sqrt{[n]} \qquad \textbf{Eq. 1}$$

Os CP possibilitaram visualizar e quantificar a contribuição das variáveis analisadas para cada neurônio resultante no mapa auto-organizável, sendo possível verificar as relações entre as respostas das várias componentes. A matriz-U possibilitou, então, a classificação dos dados relacionada ao vetor similaridade construído a partir dessas amostras.

Como resultado, foram obtidos os BMU's para cada amostra e variável analisada, bem como BMU's para esses mesmos parâmetros nas amostras com análises incompletas. A predição de valores foi determinada, portanto, a partir dos BMU's para cada amostra original, refletindo em teores sintéticos representativos para amostras onde tais teores eram originalmente desconhecidos.

É importante mencionar que para a imputação dos dados, o código SiroSOM® trabalha com a combinação de duas abordagens e variações: (i) a substituição dos valores inexistentes pelos valores BMU's; (ii) a melhora dos valores estimados mediante um processo

iterativo. Em (i), o SOM inicial é calculado e determina um conjunto inicial de valores de substituição. Já em (ii), os valores de SOM são recalculados para substituir novamente os valores não encontrados nos dados de entrada.

4.2.3. Testes de Correlação e Avaliação dos Resultados

A correlação dos resultados apresentados pelas análises clássicas e a partir da técnica SOM foi realizada mediante analises estatísticas descritivas, com a realização de tabelas e gráficos de dispersão que confrontam os valores obtidos para cada amostra estudada das variáveis aleatórias continuas: AA e SiR.

As medidas de dispersão foram feitas em torno à média, tal como a variância, a covariância, para determinar a correlação entre as variáveis calculadas e medidas simultaneamente, e o coeficiente de correlação com a finalidade de normalizar a covariância em intervalo -1 e 1. A correlação é calculada pela Eq. 2:

$$\boldsymbol{Corr[X,Y] = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}} \qquad \textbf{Eq. 2}$$

Consequentemente, em um gráfico de dispersão foi representada uma reta de regressão, a qual gera uma correlação linear, dado um conjunto de pares ordenados, para determinar uma relação funcional pelo método dos mínimos quadrados.

A correspondência dos valores obtidos pelas análises SOM e os estudos químicos foi medida pela diferença entre as médias e medianas dos teores, seguido pela porcentagem de erro relativo, conforme a Eq. 3:

$$\boldsymbol{ERP = \left|\frac{T_{SOM}-T_{LCT}}{T_{LCT}}\right| * 100} \qquad \textbf{Eq. 3}$$

Onde T_{SOM} representa os teores obtidos pela ferramenta de SOM e T_{LCT} os teores originais, obtido pelas analises de laboratório.

Além disso, foi calculado a média e mediana de cada variável com a finalidade de obter o erro porcentual das amostras de cada projeto.

É importante mencionar que, segundo a comparação dos valores, é possível medir o alcance de SOM mediante a avaliação da porcentagem de erro e correlação, para determinar a efetividade para a omissão de dados de AA e SiR em fatores de 20%, 30%, 40% e 50%.

5. RESULTADOS

5.1. Análises a partir da Técnica "Self-Organizing Maps"

Foram produzidos 12 mapas auto-organizados visualizados pela Matriz-U e cada um dos CP (variáveis) em cada etapa da exclusão das amostras dos três projetos. O número de linhas e colunas foi calculado a partir do tamanho de mapa desejado ($Size_{som}$).

Além de isso, ao fim de cada análise SOM, foi calculado o erro de quantificação final (Q_e), que representa a distancia média entre cada vetor e o seu respectivo BMU, medida da resolução do mapa. Foi calculado, ainda, o erro topográfico final (T_e), que simula a proporção de todos os vetores dos dados para os quais os principais BMU (primeiro e segundo) não são unidades adjacentes.

A tabela 2 apresenta os parâmetros preenchidos e calculados na etapa de inicialização da analises SOM para cada projeto, nas quatro fases de exclusão de dados de 20%, 30%, 40% e 50%, representados por $E_{20\%}$, $E_{30\%}$, $E_{40\%}$, $E_{50\%}$ respectivamente.

Tabela 2- Valores da etapa de inicialização de SOM

	Inicialização									
	$Size_{som}$		$E_{20\%}$		$E_{30\%}$		$E_{40\%}$		$E_{50\%}$	
	Filas	Col	Q_e	T_e	Q_e	T_e	Q_e	T_e	Q_e	T_e
Projeto A	10	14	0,341	0,190	0,300	0,190	0,296	0,228	0,274	0,271
Projeto B	8	9	0,388	0,265	0,426	0,151	0,388	0,146	0,389	0,196
Projeto C	6	7	0,304	0,0286	0,147	0	0,419	0,100	0,391	0

Em seguida, no processo de treinamento, foi selecionado o tipo de vizinho para cada vetor gaussiano. Além de isso, foi preciso definir dados robustos (*rough*) que tinham por default o raio inicial R_{i1}, raio final R_{f1} e comprimento de treinamento L_1, e os dados finos (*fine*) calculados pelo SOM, apresentados na tabela 3.

A fim de representar a estrutura e os padrões das amostras de entrada pelas similaridades, o SOM utilizou os dados de raio e comprimento inicial para cada valor de amostra de entrada.

Tabela 3- Valores robustos e finos do processo de treinamento de SOM

	Treinamento					
	R_{i1}	R_{f1}	L_1	R_{i2}	R_{f2}	L_2
Projeto A	18	5	20	5	1	400
Projeto B	13	4	20	4	1	400
Projeto C	10	3	20	3	1	400

Após a inicialização e o treinamento dos vetores, foram gerados os mapas de *Components Plots* (CP) de cada variável, e a integração dos mesmos na visualização da Matriz-U apresentados na figura 1, figura 2 e figura 3 para os projetos A, B e C, respetivamente. A escala de cores de cada uma das figuras representa a contribução das variaveis para cada umo dos mapas CP, e o nível de dissimilaridade do mapa de Matriz-U.

Em primer lugar, na figura 1, os mapas de CP apresentam alta contribuição das variáveis AA e WT, representada pela prevalência das cores quentes, a qual aumentou proporcionalmente com a exclusão de dados para AA, e permaneceu aparentemente constante para WT.

No entanto, os CP de Fe_2O_3 apresentam uma tendência contraria às de WT e AA, representada por cores frias que indicam baixa contribuição da variável, com tendência constante. No caso dos CP da SiR, o padrão de contribuição é muito menos claro. Porém, similar aos anteriores, observa-se uma baixa contribuição da variável que permaneceu relativamente constante para as quatro fases de exclusão de dados.

É assim como na Matriz-U, observa-se que a alta similaridade está, sobretudo, associada à elevada contribuição da AA e à baixa contribuição de Fe_2O_3, evidenciando correlação negativa da contribuição destas variáveis.

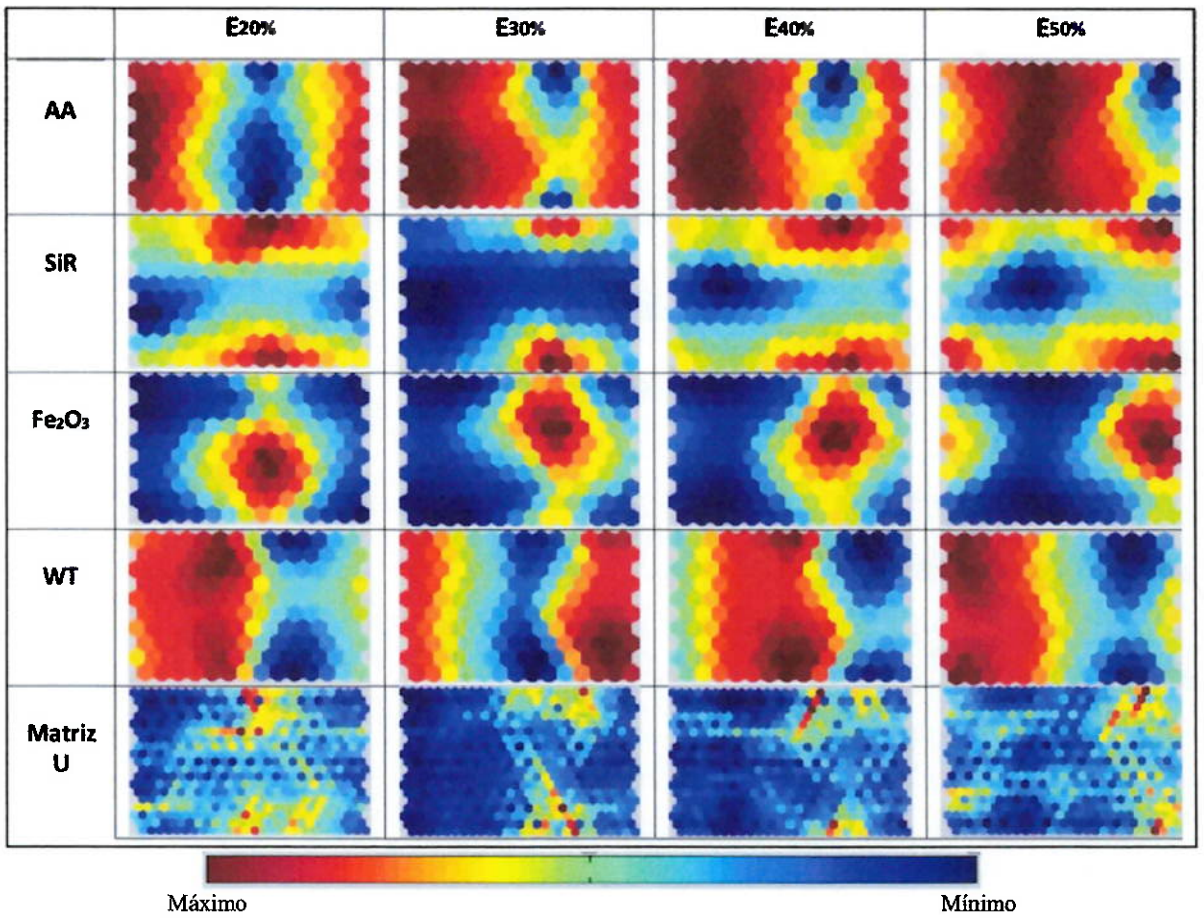


Figura 1- Mapas auto-organizados - Projeto A: CP para cada variável e Matriz-U.

Na figura 2 são apresentados os mapas auto-organizados CP para cada uma das sete variáveis do Projeto B, seguido de suas respectivas visualizações a partir da Matriz-U.

Em relação às CP, a figura 2 pode-se observar comoa AA segue o mesmo padrão de alta contribuição que $\%Al_2O_3$ Total, assim como %PF. O contrário ocorre para os CP de SiR e $\%SiO_2$ Total, $\%TiO_2$ e Fe_2O_3, que seguem o padrão de baixa contribuição dos nós para as variáveis.

Apesar da contribuição das variáveis permanecerem relativamente constante segundo sua representação de cores, as tendências e distribuições variam segundo a exclusão de dados.

A maior contribuição aos grupamentos evidentes na Matriz-U é dada pelas variáveis AA, $\%Al_2O_3$ Total e %PF. Não foram evidenciadas grandes mudanças quando comparando os resultados de menores ou maiores exclusões de dados.

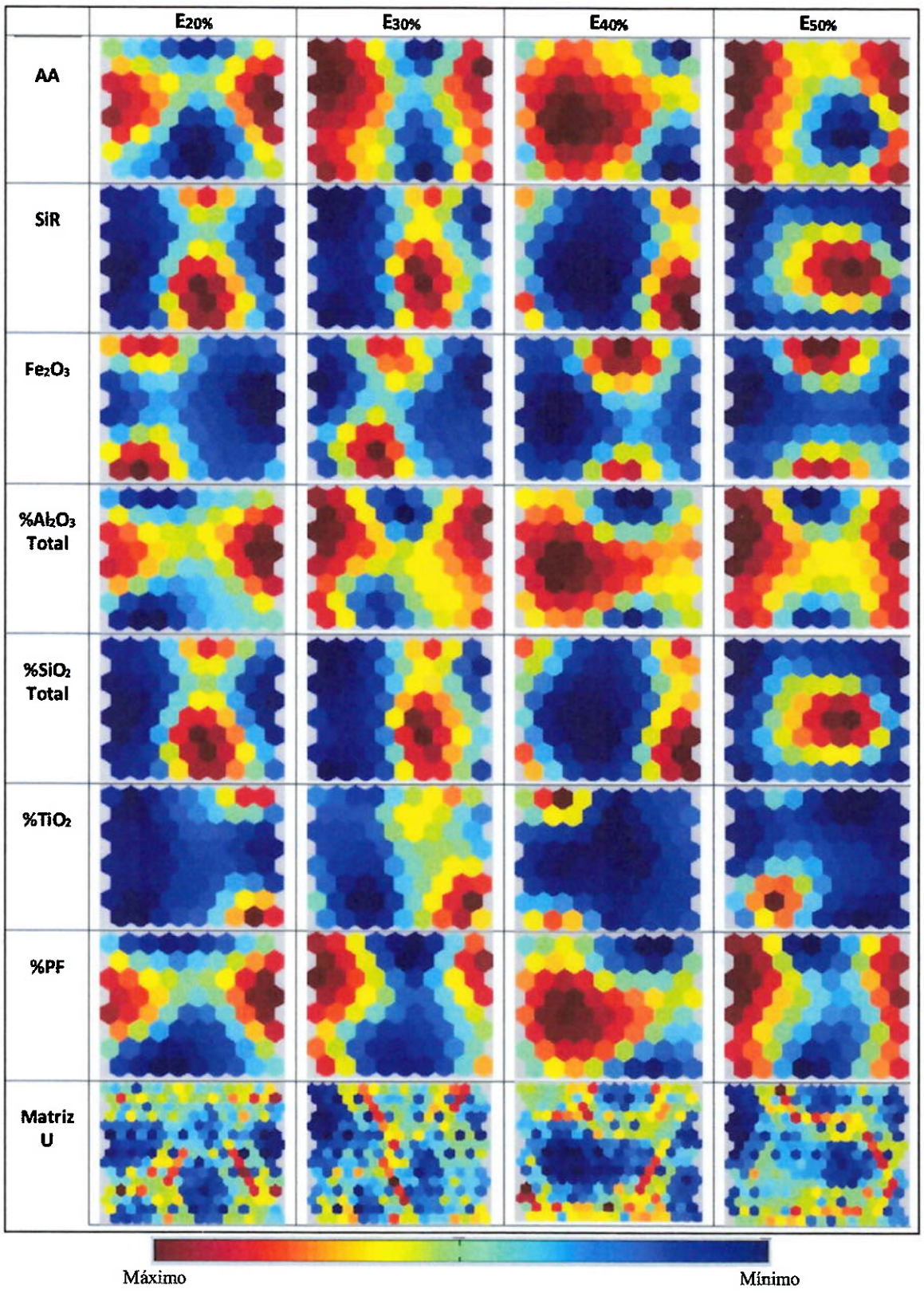


Figura 2- Mapas auto-organizados - Projeto B: CP para cada variável e Matriz-U.

Na figura 3 são mostrados os mapas auto-organizados CP das seis variáveis do Projeto C e suas respetivas visualizações na Matriz-U.

As visualizações de CP das variáveis dependentes da Alumina (AA e $\%Al_2O_3$ Total), seguem um padrão similar, independente das exclusões de dados, de médias a altas contribuições de variáveis. Em contraste, a variável Fe2O3 configura baixa contribuição em geral. Já as variáveis SiR, $\%SiO_2$ Total e $\%TiO_2$, dependentes da sílica nas amostras, seguem um padrão completamente contrário ao das variáveis da alumina, onde predominam cores frias, indicativas de contribuição media a alta.

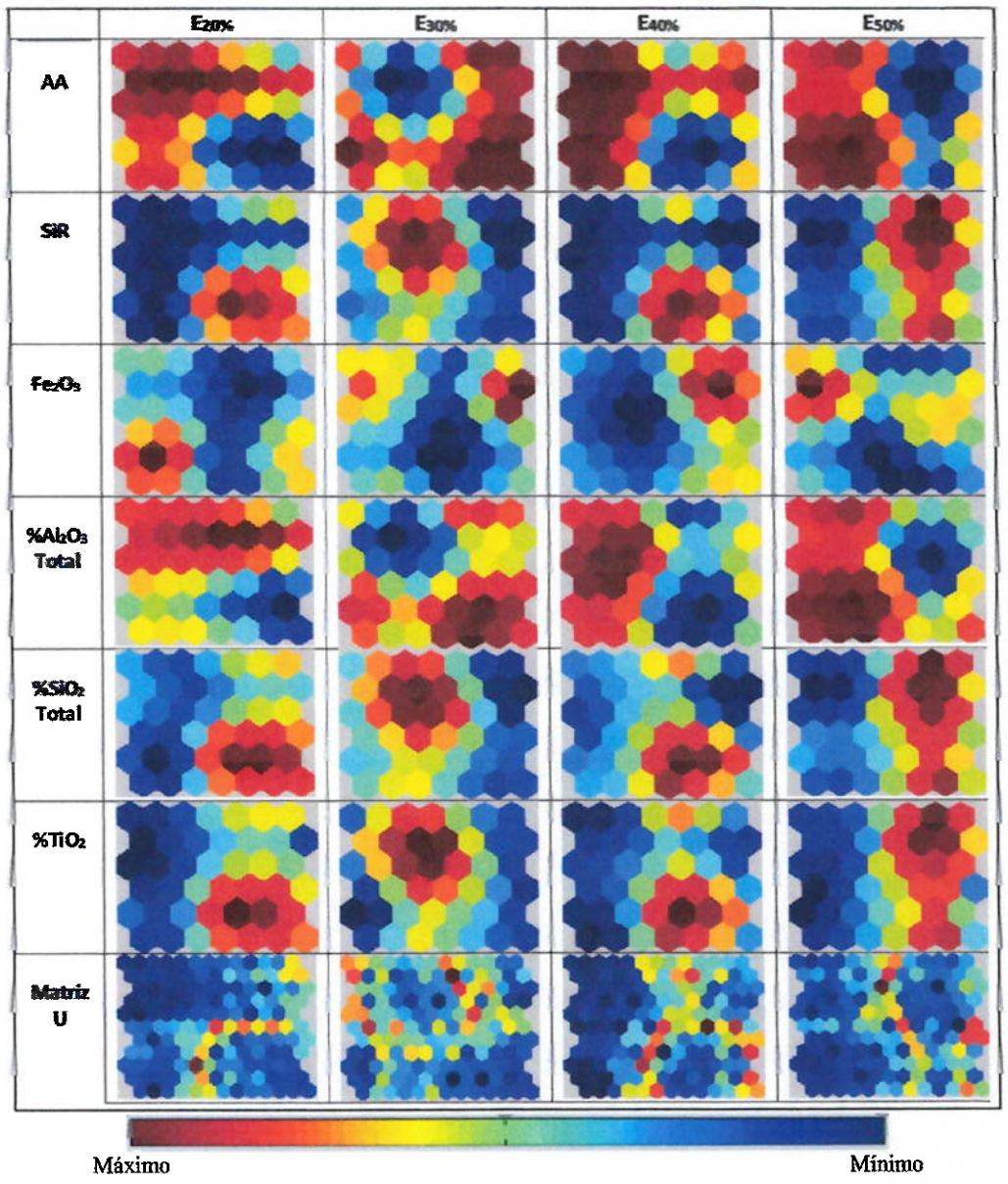


Figura 3- Mapas auto-organizados - Projeto C: CP para cada variável e Matriz-U.

Os agrupamentos de dados na Matriz-U para cada uma das exclusões não seguiram um padrão especifico. No entanto, há uma clara separação de, pelo menos, dois agrupamentos caracterizados por altas contribuições de alumina e sílica em regiões distintas.

Em termos gerais, a visualização dos mapas auto-organizados em CP demonstra a contribuição e influência de cada variável para a visualização Matriz-U, com a presença das cores quentes para representar altas contribuições, e frias no caso de baixas contribuições na predição de dados. Desse modo, embora não seja possível distinguir um padrão de comportamento segundo as porcentagens de exclusão de dados de AA e SiR nos três projetos, é possível entender que não há grandes variações na Matriz-U à medida em que aumentam as exclusões de dados. Fato que demonstra a relação direita das variáveis e seus mapas de CP para com o mapa integrado Matriz-U, pelo qual é feita a predição de dados.

5.2. Valores preditos a partir da Técnica "Self-Organizing Maps"

A técnica SOM, com o software SiroSOM® permitiu predizer 1414 pares de dados excluídos das variáveis de teores de Al_2O_3 aproveitável e SiO_2 reativa, principais elementos de controle na cadeia de produção do alumínio. Além disso, a partir dos ajustes por BMU, foram obtidos novos valores para cada uma das amostras nas variáveis de entrada.

5.3. Correlação e Avaliação dos Resultados

Com a finalidade de avaliar e simplificar a visualização dos resultados obtidos pelo SOM, foi calculada a média (M_t) e mediana do teor (M_{et}) de AA e SiR dos dados originais, bem como os obtidos pelo BMU gerados pelas análises em SOM.

Este procedimento foi realizado para cada projeto e suas diferentes derivações com porcentagens de exclusão de dados. Os resultados são apresentados nas tabelas 4, 5, 6 e 7 com a finalidade de comparar no só a influencia da porcentagem excluída, mas também as diferenças quanto à utilização das variáveis e amostras de cada projeto.

Em termos gerais, pode-se notar que os coeficientes de correlação tanto da AA como de SiR de todos os projetos é efetiva, ou seja, mantém uma relação positiva, que indica proporcionalidade direita (Tabelas 4, 5, 6 e 7. Apêndices A e B). Os resultados são coerentes, portanto, levando em conta de que tratam da comparação de uma mesma variável.

Tabela 4- Análises estadístico de AA e SiR. Projetos A, B, C. Exclusão 20%.

	Projeto A				Projeto B				Projeto C			
	AA	B_AA	SiR	B_SiR	AA	B_AA	SiR	B_SiR	AA	B_AA	SiR	B_SiR
M_t	36,92	37,09	6,24	5,72	32,98	31,83	10,38	10,34	37,94	37,96	10,37	10,43
M_{et}	40,90	40,62	5,17	5,23	36,06	5,42	33,97	5,72	48,90	48,43	4,40	3,95
Max_t	57,9	52,2	47,9	13,5	53,7	50,4	30,9	26,9	52,6	50,6	31,1	26,4
Min_t	3,0	10,9	1,2	2,1	5,7	8,0	2,4	4,0	9,2	10,9	1,5	2,9
ERP(%)	0,69		1,09		565,40		494,40		0,97		11,34	
Corr	0,97		0,74		0,99		0,99		1,00		0,99	

Tabela 5- Análises estadístico de AA e SiR. Projetos A, B, C. Exclusão 30%.

	Projeto A				Projeto B				Projeto C			
	AA	B_AA	SiR	B_SiR	AA	B_AA	SiR	B_SiR	AA	B_AA	SiR	B_SiR
M_t	36,92	35,76	6,24	6,31	32,98	31,80	10,38	10,34	37,94	35,39	10,37	8,87
M_{et}	40,90	38,94	5,17	5,47	36,06	32,49	5,42	5,68	48,90	42,14	4,40	4,12
Max_t	57,9	51,8	47,9	18,2	53,7	49,9	30,9	26,7	52,6	49,6	31,1	21,9
Min_t	3,0	6,1	1,2	2,4	5,7	8,1	2,4	4,3	9,2	10,4	1,5	2,2
ERP(%)	5,03		5,50		10,97		4,61		16,04		6,80	
Corr	0,95		0,73		0,99		0,99		0,98		0,99	

Tabela 6- Análises estadístico de AA e SiR. Projetos A, B, C. Exclusão 40%.

	Projeto A				Projeto B				Projeto C			
	AA	B_AA	SiR	B_SiR	AA	B_AA	SiR	B_SiR	AA	B_AA	SiR	B_SiR
M_t	36,92	35,87	6,24	5,67	32,98	31,83	10,38	10,27	37,94	35,87	10,37	10,26
M_{et}	40,90	38,85	5,17	5,33	36,06	34,64	5,42	6,10	48,90	48,79	4,40	4,59
Max_t	57,9	51,3	47,9	12,9	53,7	50,1	30,9	26,6	52,6	49,9	31,1	26,3
Min_t	3,0	6,4	1,2	2,2	5,7	7,9	2,4	4,2	9,2	10,7	1,5	3,7
ERP(%)	5,28		2,98		4,09		11,18		0,22		4,03	
Corr	0,95		0,67		0,99		0,99		0,98		0,99	

Tabela 7- Análises estadístico de AA e SiR. Projetos A, B, C. Exclusão 50%.

	Projeto A				Projeto B				Projeto C			
	AA	B_AA	SiR	B_SiR	AA	B_AA	SiR	B_SiR	AA	B_AA	SiR	B_SiR
M_t	36,92	35,52	6,24	5,74	32,98	31,38	10,38	10,58	37,94	38,15	10,37	8,89
M_{et}	40,90	38,10	5,17	5,44	36,06	34,71	5,42	6,30	48,90	46,74	4,40	3,71
Max_t	57,9	50,7	47,9	13,2	53,7	49,8	30,9	26,8	52,6	49,5	31,1	22,5
Min_t	3,0	7,0	1,2	2,3	5,7	8,2	2,4	4,2	9,2	12,0	1,5	2,6
ERP(%)	7,36		4,95		3,90		13,92		4,63		18,66	
Corr	0,92		0,60		0,99		0,98		0,99		0,98	

Além de isso, é possível estabelecer um faixa de correlação entre 0,98 e 1 da AA nos projetos B e C. No entanto, para o projeto A, a AA apresenta valor mínimo de 0,92 e máximo de 0,97, o que indica maior variância entre AA inicial e a calculada pelo SOM. Em geral pode-se observar que a correlação diminui na medida em que se aumenta a exclusão para os três projetos, apesar da faixa de correlação entre 0,92 e 1, que sugere altos valores.

Os valores de SiR apresentam também uma correlação inversamente proporcional à exclusão das amostras. Porém, é menor em relação à correlação da AA, o qual é mais evidente para o projeto A, com uma faixa de valores que varia entre 0,60 e 0,74. Para os projetos B e C a correlação de SiR se mantém alta, com variação entre 0,98 e 0,99.

É importante mencionar que os erros porcentuais foram calculados segundo a mediana, devido a que esta é menos sensível a flutuações nos valores médios da variável e é mais representativa para populações heterogêneas, tal como foram os grupos das variáveis originais e calculadas [29]. Os valores máximos e mínimos de teores originais e calculados pelas analises SOM para cada projeto são ilustrados nos Apêndices C, D, E, F, G e H.

6. DISCUSSÃO

Esta pesquisa, que procura um melhor alcance na predição e interpretação de dados geoquímicos, parte das limitações em relação à visualização adequada para os conjuntos de dados diversos e de alta dimensionalidade, caracterizada pelas múltiplas variáveis. Assim, a abordagem desta pesquisa lançou mão da técnica SOM, o que permitiu gerar vetores decompostos, analisados para extrair a importância relativa de cada um dos componentes durante a classificação. Tal abordagem favoreceu uma visão sobre as relações complexas em conjuntos de dados de alta dimensionalidade, como é o caso das análises geoquímicas. As análises SOM favorecem, portanto, a preservação das relações topológicas e, ao mesmo o tempo, a produção de um modelo estatístico decorrente do conjunto de dados [9].

As análises estatísticas dos resultados obtidos no projeto A resultaram em alta correlação de AA. Porém, com maior variância em relação aos projetos B e C. Já em relação aos valores de correlação para a SiR, os resultados obtidos no projeto A apresentaram resultados inferiores tanto em comparação com os projetos B e C, quanto com os resultados obtidos para a AA. Uma questão evidente é relativa ao maior número de amostras do projeto.Essas amostras não estão completamente relacionadas em regiões de origem. Além

disso, observaram-se grandes discrepâncias entre os teores dos 10 diferentes grupos de amostras, estudados como um mesmo conjunto. Isso provavelmente ocasionou maior variância e incerteza na relação das amostras de uma variável específica.

Além disso, no projeto A o máximo teor da SiR dos dados originais não é um valor frequente (não é representativo) e, por vezes, apresenta teores muito maiores à mediana. Por essa razão, pode ser considerada a possibilidade de erros analíticos em determinadas amostras. No entanto, os teores máximos calculados pelas analises SOM não foram diretamente influenciadas por esses valores. A saber, os teores calculados pelo SOM seguiram o padrão dos demais dados da variável SiR, tal como pode ser observado no apêndice D.

Em relação ao projeto B, este apresentou os melhores resultados relativos à predição dos dados, quando comparado aos projetos A e C. É notável que a maior variabilidade e dissimilaridades nos mapas de CP e, consequentemente, na Matriz-U integrados induziram à possibilidade de predição de maneira mais eficientemente. Estes resultados refletem o produto da influencia de uma maior quantidade de variáveis, sem importar a diversidade de origem das amostras de bauxitas analisadas.

Quanto ao projeto C, tornou-se clara a alta correlação tanto para AA como para SiR. Porém, o estudo foi feito com poucos dados, o que refletiu em dimensões pequenas para o tamanho do mapa auto-organizado. A saber, um mapa auto-organizado com baixa densidade de neurônios dificulta a ótima interpretação visual, e proporciona em incremento do erro associado. Além de isso, o alto valor de correlação pode refletir em baixa significância, dado o restrito número de amostras analisadas em C.

Em termos gerais, os resultados apresentaram alta correlação de valores entre as variáveis medidas em laboratório e aquelas preditas pelo SOM. No entanto, em amostras de bauxitas originárias de múltiplas fontes, foi notável que predição de dados para AA teve maior correlação com os resultados originais que as predições obtidas para a SiR. Isso pode ser explicado pela influência de outros parâmetros ou pela ausência de variáveis relacionadas ou dependentes entre si, que não estavam presentes nas análises do projeto que envolveu bauxitas provenientes de regiões diversas (Projeto A).

Como sugestões alternativas, a análise química total realizada mediante o método de absorção atómica poderia aportar variáveis tais como: porcentagem de óxidos de cálcio

(% CaO); óxido de magnésio (% MgO); e óxido de potássio(% K_2O). A inclusão destas variáveis às análises SOM seriam parâmetros influentes.

A obtenção de melhores respostas da Matriz-U poderá ser obtida a partir de análises com alta densidade de dados inter-relacionados, bem como com a maior quantidade de variáveis possíveis. A geração dos mapas CP de cada variável é integrada no mapa auto-organizado 2D (neste estudo evidenciado pela Matriz-U). Porém, a influência das diferentes naturezas e dependências das variáveis afetam a distribuição uniforme e a identificação de agrupamentos neste mapa. Isso é considerado como um fator influente na melhor predição de dados, ou seja, a baixa correlação entre as distintas variáveis originais resulta numa ótima predição de dados desconhecidos.

7. CONCLUSÕES

A predição de 1414 pares de valores para AA e SiR, mediante o uso da técnica não-supervisionada SOM, permitiu constatar a eficiência da técnica como ferramenta complementar à geração de dados analíticos. Assim, o SOM passa a ser explorado não apenas como ferramenta de classificação, integração e interpretação de dados multivariados, como a maioria dos estudos atuais o reconhecem, mas como uma ferramenta capaz de predizer teores analíticos.

Baseado nas analises estatísticas desenvolvidas, a alta correlação entre os valores originais medidos pelas analises químicas em laboratório e aquelas preditas por SOM permitiu definir o a técnica como efetiva para prever dados com até 50% de ausência de valores em até duas variáveis simultâneas entre outras variáveis.

Em relação à influencia dos parâmetros e variáveis utilizadas neste estudo, a técnica se demonstrou mais eficiente quando utilizada em amostras originárias de fontes próximas. Desse modo, as análises proporcionar o uso mais adequado da ferramenta SOM, implicando em menores erros de amostragem.

Quanto maior for a quantidade de variáveis analíticas de entrada, menores serão os erros associados à predição de dados com o SOM. Desse modo, a realização de mais análises químicas convencionais poderia gerar maior quantidade de variáveis, as qual podem proporcionar melhores resultados para as predições, especificamente no caso da SiR.

Os estudos ora apresentados foram desenvolvidos em variáveis de interesse específicos (AA e SiR). No entanto, outros estudos podem ser desenvolvidos futuramente no sentido de explorar a predição em maiores percentagens de exclusão de valores (acima de 50%), bem como em maiores quantidades de variáveis.

8. REFERENCIAS

[1] S. J. FRASER, AND B. L. DICKSON, **A New Method for Data Integration and Integrated Data Interpretation: Self-Organising Maps**. Proceedings of Exploration 07: Fifth Decennial International Conference of Mineral Exploration, 2007.

[2] L. SEVILLANO. **Quantificação de minerais de Bauxita pordifração de Raios X e sua correlação comanálise química.** Escola Politécnica, Universidade de São Paulo, São Paulo, 2010.

[3] C. I. E. N. **Código Geológico de Venezuela**. 1997.

[4] ABAL.**Associação Brasileira de Alumínio-ABAL**, São Paulo. , 2007.

[5] J. L. ANTONIASSI. **A difração de raios X com o método de Rietveld aplicada a bauxitasde Porto Trombetas, PA**. Escola Politécnica Universidade de São Paulo, São Paulo, 2010.

[6] T. KOHONEN. **Self-Organizing Maps**. Third ed., Helsinki University of Technology Neural Networks Research Centre, Finland., 2001.

[7] J. D. WALLS, M. T. TANER, G. TAYLOR et. al. **Seismic Reservoir Characterization of a Mid-continent Fluvial System Using Rock Physics, poststack Seismic Attributes And Neural Networks.** 2000.

[8] U. STRECKER, AND R. UDEN. **Data mining of 3D poststack seismic attribute volumes using Kohonen self-organizing maps.** The Leading Edge, vol. 21, no. 10, pp. 1032-1037, October 1, 2002.

[9] B. S. PENN. **Using self-organizing maps to visualize high-dimensional data**. Comput. Geosci., vol. 31, no. 5, 2005.

[10] C. D. C. CARNEIRO, S. J. FRASER, A. P. CRÓSTA et al. **Semiautomated geologic mapping using self-organizing maps and airborne geophysics in the Brazilian Amazon**. Geophysics, vol. 77, no. 4, pp. K17-K24, July 1, 2012.

[11] M. J. CRACKNELL, A. M. READING, AND P. DE CARITAT. **Multiple influences on regolith characteristics from continental-scale geophysical and mineralogical remote sensing data using Self-Organizing Maps**. Remote Sensing of Environment, vol. 165, 2015.

[12] J. MACGREGOR. **Neural Modeling**, 1977.

[13] L. D. HARMON, AND E. R. LEWIS. **Neural modeling**. 1966.

[14] J. MOHD ALI, M. A. HUSSAIN, M. O. TADE et al. **Artificial Intelligence techniques applied as estimator in chemical process systems – A literature survey.** Expert Systems with Applications, vol. 42, no. 14, 2015.

[15] T. KOHONEN. **The self-organizing map.** Neurocomputing, vol. 21, no. 1–3, 1998.

[16] E. OJA. **Principal components, minor components, and linear neural networks.** Neural Networks, vol. 5, no. 6, 1992.

[17] T. KOHONEN, S. KASKI, AND H. LAPPALAINEN. **Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM.** 1997.

[18] Z. HUICHENG, G. LEFEBVRE, AND C. LAURENT. **Fast-Learning Adaptive-Subspace Self-Organizing Map: An Application to Saliency-Based Invariant Image Feature Construction.** Neural Networks, IEEE Transactions on, vol. 19, no. 5, 2008.

[19] K. HORIO, AND T. YAMAKAWA. **Feedback adaptive subspace self-organizing map for robust spatio-temporal pattern classification.** International Congress Series, vol. 1269, 2004.

[20] A. ULTSCH. **U-matrix: A tool to visualize clusters in high dimensionaldata**. University of Marburg, Department of Computer Science, TechnicalReport, 36, 12., 1993.

[21] A. ULTSCH, AND C. VETTER. **Self-organising feature maps versus statistical clustering a benchmark.** Technical Report No. 9, Dept. of Mathematics and Computer Science, University of Marburg, Germany., 1994.

[22] M. SIPONEN, J. VESANTO, O. SIMULa et al.**An approach to automated interpretation of SOM**. Advances in Self-Organising Maps, Springer London, 2001.

[23] T. AOKI, K. OTA, K. KURATA et al. **Ordering process of self-organizing maps improved by asymmetric neighborhood function.**Cognitive Neurodynamics,vol.3,no.1 2009.

[24] S. RODRÍGUEZ. **Recursos Minerales de Venezuela.** Boletín del Ministerio de Energía y Minas, 1986.

[25] KAUFMANS W. **X-Ray Fluorescence.**Journal of the Röntgen Society, vol. 10, no. 38, 1914.

[26] R. HENDERSON. **The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules.**Quarterly Reviews of Biophysics, vol. 28, no. 02, 1995.

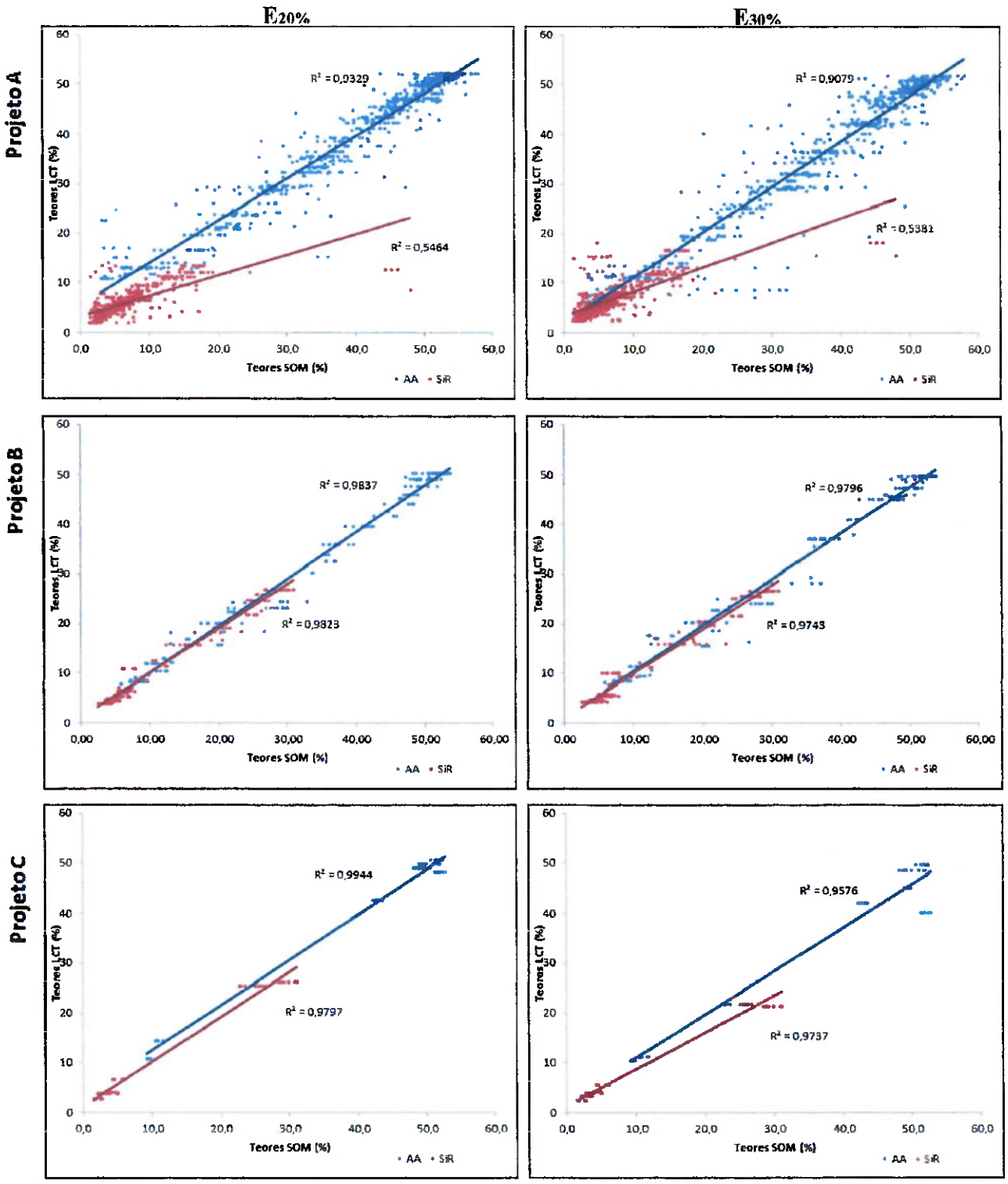
[27] H. S. MALEK M.A., S. S.M., AND M. I. **Imputation of time series data via Kohonen self-organizing maps in the presence of missing data.** Introduction to Geophysical Prospecting. 4th. Ed. McGraw-Hill. no. Engineering and Technology 41:501–506.Dobrin M.B., Savit C.H. 1988.2008.

[28] J. VESANTO, J. HIMBERG, E. ALHONIEMI et al. **SOM Toolbox for Matlab 5.**, H. U. o. T. Neural Networks Research Centre, Helsinki, Finland, 2000.

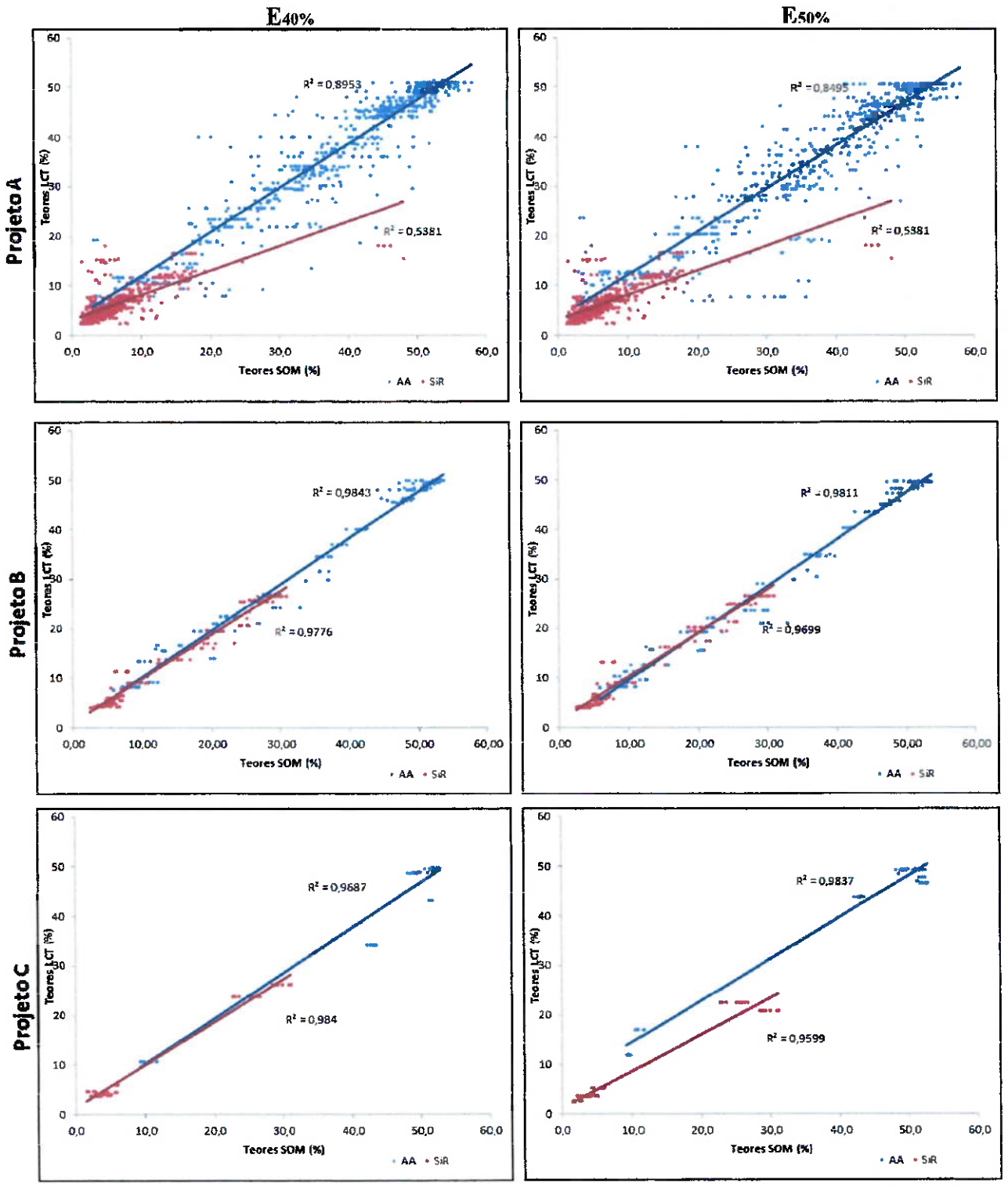
[29] J. C. DAVIS. **Statistics and Data Analysis in Geology**. John Wiley Sons, Inc., 1990.

9. APÊNDICE

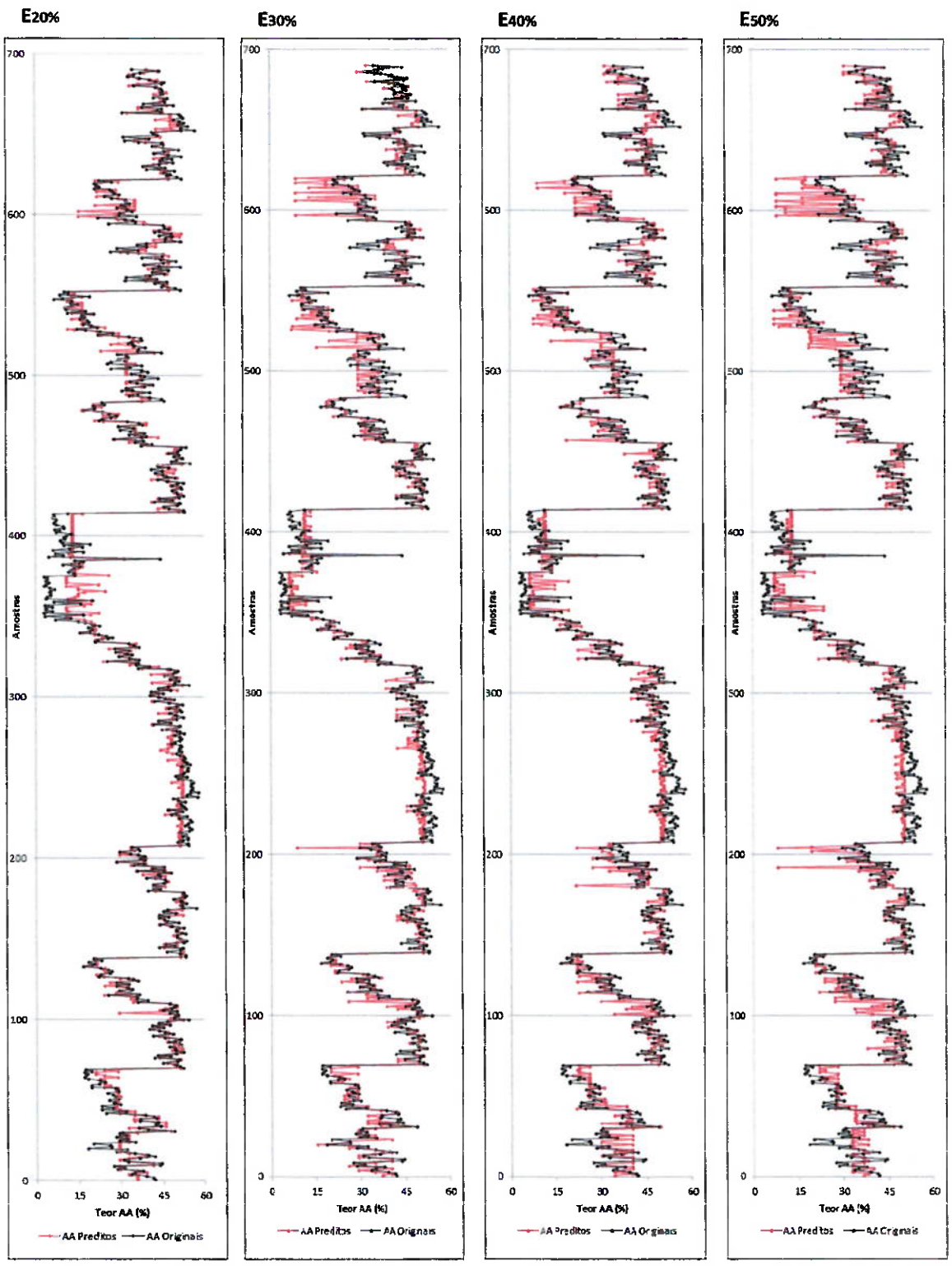
9.1. APÊNDICE A- Correlaçao de AA e SIR. E20% e E30%



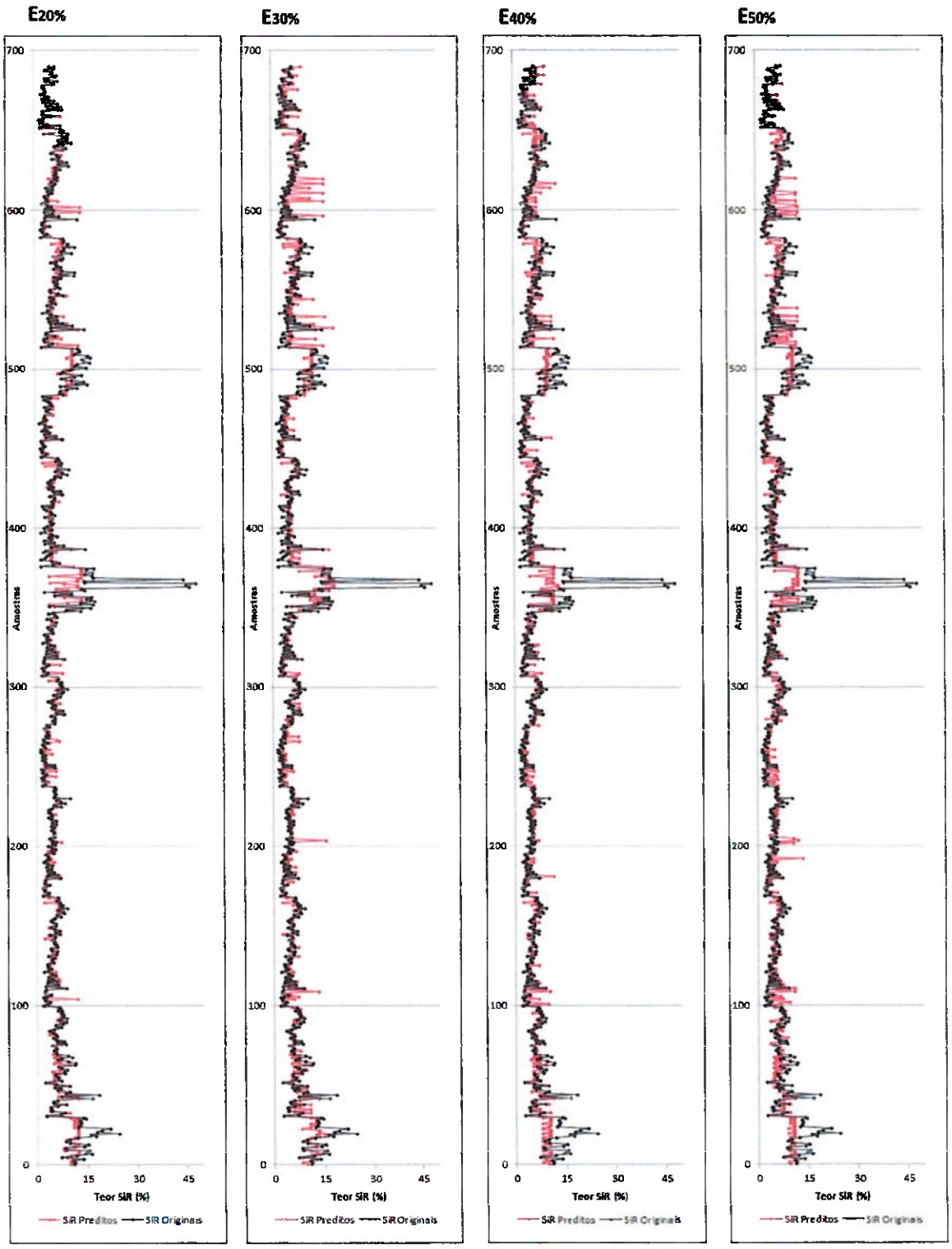
9.2. APÊNDICE B- Correlação de AA e SiR. E40% e E50%



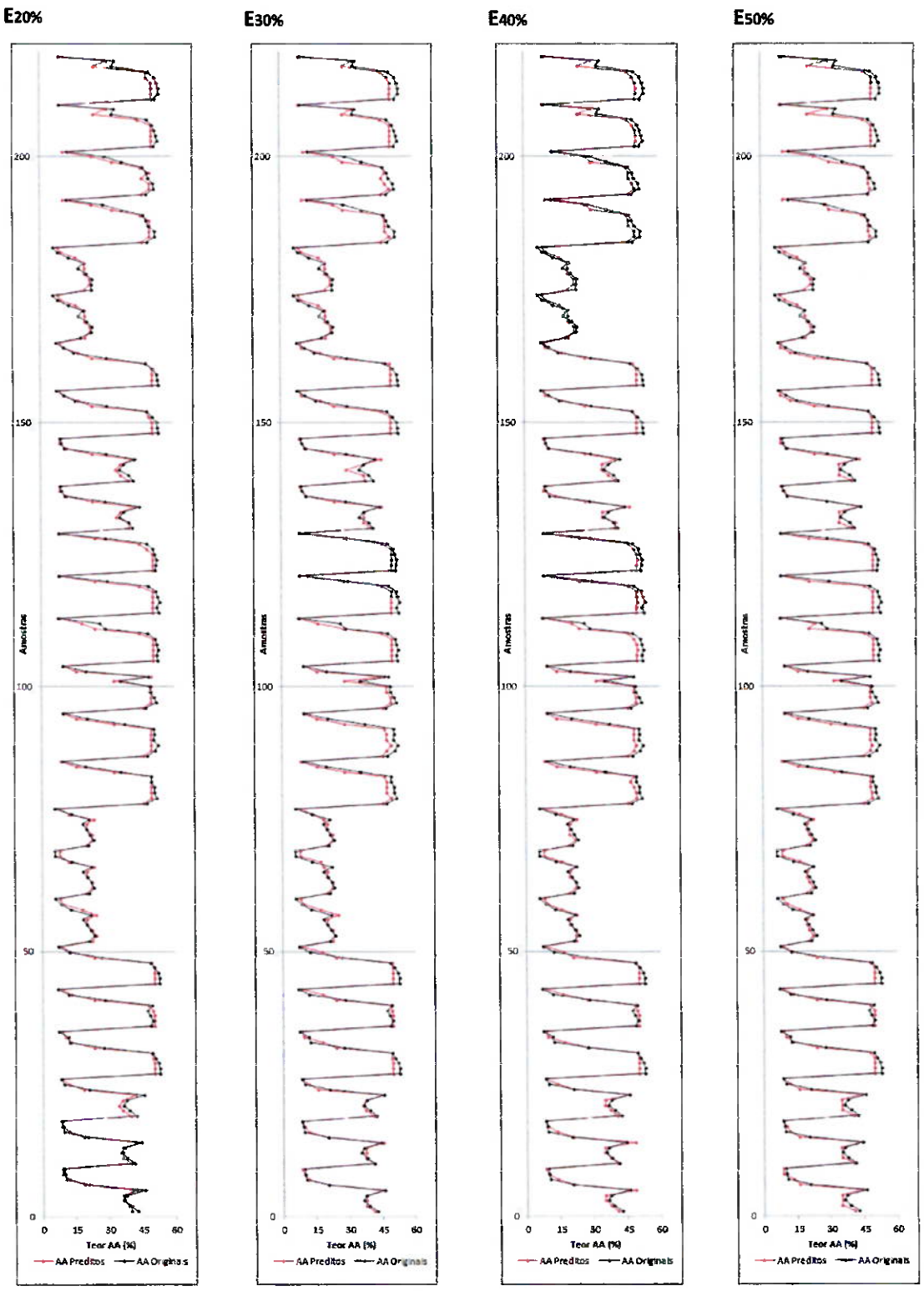
9.3. APÊNDICE C- Comparação de Teores Originais e Preditos de AA. Projeto A



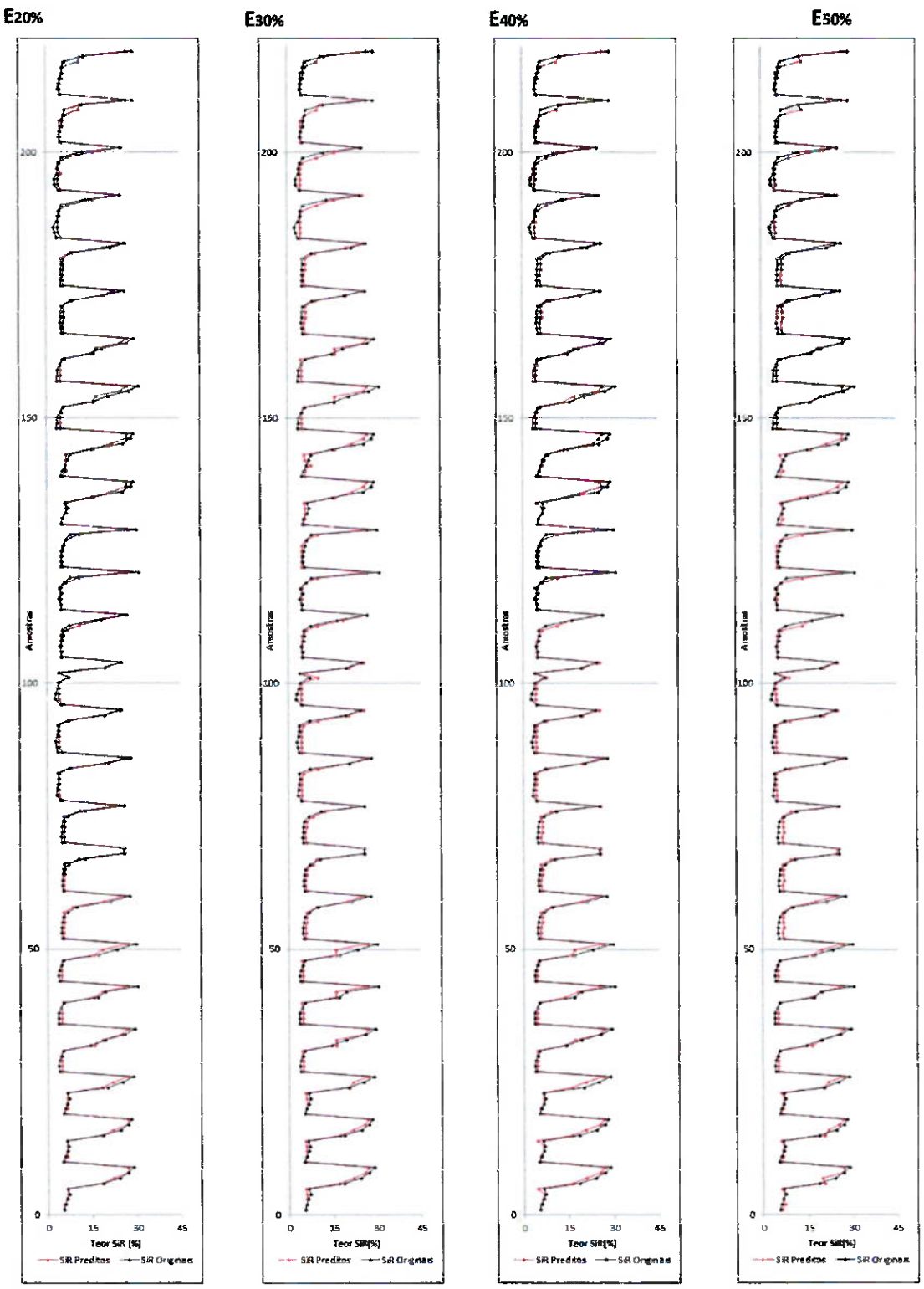
9.4. APÊNDICE D- Comparação de Teores Originais e Preditos de SiR. Projeto A



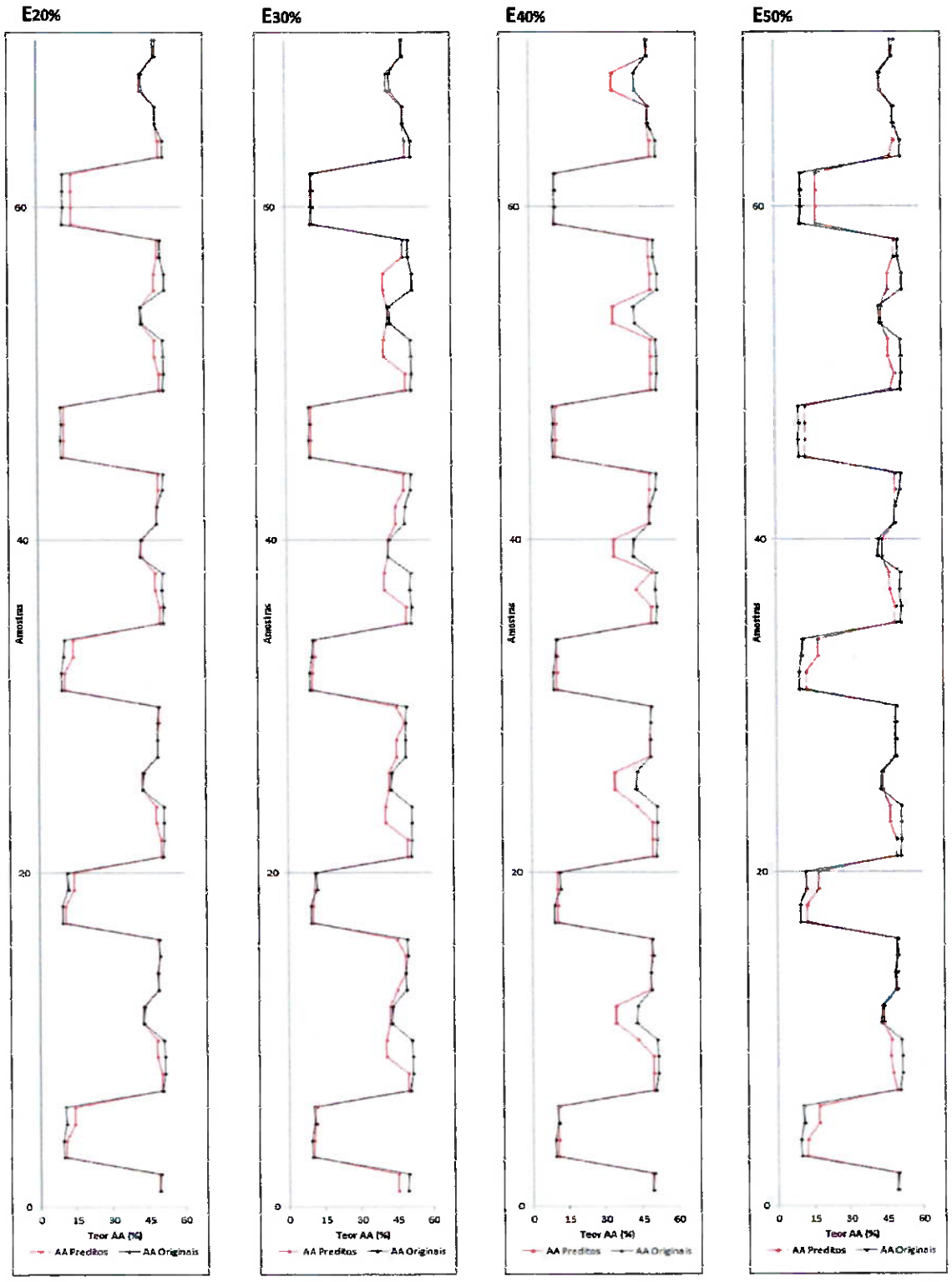
9.5. APÊNDICE E-Comparação de Teores Originais e Preditos de AA. Projeto B



9.6. APÊNDICE F-Comparação de Teores Originais e Preditos de SiR. Projeto B



9.7. APÊNDICE G-Comparação de Teores Originais e Preditos de AA. Projeto C



9.8. APÊNDICE H-Comparação de Teores Originais e Preditos de SiR. Projeto C