

UNIVERSIDADE ESTADUAL DE PONTA GROSSA  
MESTRADO EM COMPUTAÇÃO APLICADA

KARINE SATO DA SILVA

**USO DA MINERAÇÃO DE DADOS PARA EXTRAÇÃO DE CONHECIMENTO  
AGRONÔMICO ENVOLVENDO O USO DE GESSO AGRÍCOLA**

PONTA GROSSA  
2013

KARINE SATO DA SILVA

**USO DA MINERAÇÃO DE DADOS PARA EXTRAÇÃO DE CONHECIMENTO  
AGRONÔMICO ENVOLVENDO O USO DE GESSO AGRÍCOLA**

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada – Área de concentração Computação para Tecnologias em Agricultura – Universidade Estadual de Ponta Grossa como requisito para obtenção do título de Mestre em Computação Aplicada.

Orientador: Prof<sup>ª</sup> Dra. Elaine Margarete Guimarães  
Co-orientador: Prof<sup>º</sup> Dr. Eduardo Fávero Caires

PONTA GROSSA

2013

Ficha Catalográfica Elaborada pelo Setor Tratamento da Informação Bicen/UEPG

S586

Silva, Karine Sato da

Uso da mineração de dados para extração de conhecimento agrônômico envolvendo o uso de gesso agrícola/ Karine Sato da Silva. Ponta Grossa, 2013. 93f.

Dissertação (Mestrado em Computação Aplicada - área de concentração Computação Aplicada a Agricultura), Universidade Estadual de Ponta Grossa. Orientadora: Profª Drª Elaine Margarete Guimarães. Co-orientador: Prof Dr Eduardo Fávero Caires.

1. Gesso agrícola. 2. Mineração de dados. 3. Sistema plantio direto 4. Seleção de atributos. 5. *Mega-trend-diffusion*. I. Guimarães, Elaine Margarete. II. Caires, Eduardo Fávero. III. Universidade Estadual de Ponta Grossa. Mestrado em Computação Aplicada. IV. T.

CDD: 630.2

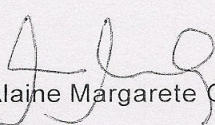
## TERMO DE APROVAÇÃO

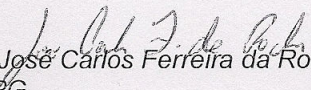
Karine Sato da Silva

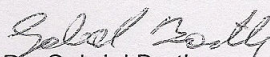
“USO DA MINERAÇÃO DE DADOS PARA EXTRAÇÃO DE CONHECIMENTO  
AGRONÔMICO ENVOLVENDO O USO DE GESSO AGRÍCOLA”

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, pela seguinte banca examinadora:

Orientador:

  
Dr.ª - Elaine Margarete Guimarães  
UEPG

  
Dr. José Carlos Ferreira da Rocha  
UEPG

  
Dr. Gabriel Barth  
FUNDAÇÃO ABC

Ponta Grossa, 01 de março de 2013.

Dedico este trabalho a Eliane, Marcos, Camille e Leopoldo.

meus grandes amores

Dedico este trabalho a Eliane, Marcos, Camile e Leopoldo,  
meus grandes amores.

## **AGRADECIMENTOS**

Agradeço a DEUS por me guiar em todos os momentos, me abençoando e protegendo todos os dias.

Agradeço aos meus pais, Marcos e Eliane, pelo incentivo a estudar, pelos conselhos, e por estarem presentes em todos os momentos da minha vida. Agradeço a minha irmãzinha Camile, por ser minha fonte de alegria.

Agradeço a minha família inteira, meus avós, Luiza e Mário, e, Ana Cacilda e Sebastião (em memória), meus tios, Willian e Kayo, e, Delvana e Ariangelo, meus primos, Ariangelo, Caio, Flávia e Kaori, pelo apoio e por estarem sempre presentes.

Agradeço ao Leopoldo, meu namorado e melhor amigo, pela sua compreensão e incentivo.

Agradeço a Adriana, Lourenço, Leonardo, Gustavo, Ivo e Fátima, pelo carinho e consideração.

Agradeço aos meus grandes amigos, que estão sempre presentes, e me motivando em cada momento, Letícia, Aniele, Sabrina, Rúbia e Guto.

Agradeço a minha orientadora e amiga Prof<sup>a</sup> Dra. Alaine, a quem tenho muito carinho, que me transmitiu o conhecimento necessário para que fosse possível a realização deste trabalho.

Agradeço ao meu co-orientador Prof<sup>o</sup> Dr. Eduardo, pela disposição em ajudar no desenvolvimento deste trabalho e pelo fornecimento das bases de dados utilizadas.

Agradeço a todos os professores e amigos da UEPG que de alguma forma contribuíram para minha formação acadêmica, em especial, Prof<sup>o</sup> Dr. José Carlos, Thalita e Juscelino que me auxiliaram em muitos momentos no decorrer desses dois anos de estudo.

Agradeço a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudos.

## RESUMO

A acidez do subsolo é prejudicial ao crescimento das raízes das plantas e, conseqüentemente, afeta a produtividade agrícola. Em áreas manejadas no sistema plantio direto (SPD), os efeitos tóxicos causados pelos teores elevados de Al e Mn, decorrentes da acidez do solo, são corrigidos pela calagem superficial. Essa prática melhora a acidez das camadas superficiais, mas não apresenta grande eficiência na correção da acidez em camadas mais profundas do solo. O gesso agrícola ( $\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$ ) é um insumo que pode auxiliar na melhoria das condições químicas do subsolo, pois além de ser fonte de Ca e S, também é capaz de transportar nutrientes catiônicos para as camadas subsuperficiais e reduzir a atividade do Al. Acontece, porém, que ainda existem dúvidas sobre quais as situações em que podem ser esperados efeitos benéficos do uso de gesso agrícola e também quanto à dose que deve ser aplicada para alcançar tais efeitos. Uma forma possível para auxiliar na compreensão dessas dúvidas é a utilização da Mineração de Dados (MD). Entretanto, as bases de dados agronômicos normalmente envolve um número limitado de registros, o que dificulta o emprego da MD. Em decorrência disso, estudou-se, além da utilização da MD, uma nova área de pesquisa que envolve a MD em pequenas bases. Com isso, os objetivos do presente trabalho foram: (i) obter melhor compreensão dos efeitos da aplicação do gesso nos atributos químicos de solos manejados no SPD, (ii) identificar os atributos químicos do solo que apresentam correlações mais estreitas com a estimativa da necessidade de gesso usando técnicas de seleção na etapa de pré-processamento, e (iii) definir modelos para estimativa de necessidade de gesso agrícola em solos sob SPD. A base de dados utilizada nesse estudo foi obtida de três áreas distintas da região dos Campos Gerais do Paraná, contendo atributos químicos do solo em diferentes épocas provenientes de experimentos sob SPD que receberam doses crescentes de gesso agrícola na superfície. Utilizaram-se técnicas de Análise de Componentes Principais (ACP) baseada nos critérios B2 e B4 e, ainda, a técnica de ACP Supervisionado. Quanto às técnicas de pré-processamento, implementou-se também uma matriz de covariância que assume a independência marginal entre os atributos da base no seu cálculo e utiliza os critérios B2 e B4 para a seleção dos atributos. Para ampliação da base de dados, além da técnica SMOTE, foi empregado o método *mega-trend-diffusion* (MTD). O algoritmo *M5Rules* foi utilizado para encontrar modelos de estimativa da necessidade de gesso agrícola. Os resultados mostraram que o tempo decorrido após a aplicação de gesso (época), a saturação por Ca e a saturação por Mg na capacidade de troca de cátions efetiva (CTCe) do solo foram os atributos que apresentaram as correlações mais estreitas com a estimativa da dose de gesso. O trabalho identificou quatro possíveis modelos para a estimativa da necessidade de gesso, mostrando que o algoritmo *M5Rules* foi eficiente para tal predição. O emprego do método MTD apresentou resultados positivos porque aumentou o coeficiente de correlação e diminuiu o erro médio absoluto.

**Palavras-chave:** gesso agrícola, mineração de dados, sistema plantio direto, seleção de atributos, *mega-trend-diffusion*

## ABSTRACT

The subsoil acidity is harmful towards the growing of plants roots and, consequently, affects the agricultural productivity. In handled areas in the no-till cropping system (SPD), the toxic effects caused by high levels of Al and Mn, due to the soil acidity, are corrected by the superficial liming. This technique improves the acidity of the superficial layers, but it presents no great efficiency in the acidity correction of deeper layers of the soil. The agricultural gypsum ( $\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$ ) is an input that might help in the improvement of the subsoil's chemical conditions, because besides of being a Ca and S source, it is also able to transport cationic nutrients to the sub superficial layers and reduce the Al activity. It happens, however, that there are questions about which situations may be expected beneficial effects regarding the agricultural gypsum use, and as for the amount that should be applied to reach such effects. A possible form to assist the comprehension of these questions is with the Data Mining (MD) utilization. However, the agronomic databases usually involve a limited number of registers, which difficult the MD use. As a result, this study addresses, beyond the MD utilization, a new research area involving MD in small databases. Therewith, the goals of this work were: (i) obtaining a better comprehension of the gypsum application effects in the chemical attributes of handled soil in SPD, (ii) identifying the chemical attributes of the soil that present narrower correlations with the estimation of the need of gypsum using selection techniques at the pre-processing stage, and (iii) defining models to the estimation of the need of agricultural gypsum in soils under SPD. The database used in this study was obtained from three distinct areas of the region of Campos Gerais do Parana, containing chemical attributes of the soil in different epochs coming from SPD experiments, which received increasing doses of agricultural gypsum on their surfaces. It was used Principal Component Analysis techniques based on B2 and B4 criteria, and also the Supervised ACP technique. Still regarding the pre-processing techniques, it was implemented a covariance matrix that assumes the marginal independence between the base attributes in their calculus and utilizes the B2 and B4 criteria for the attributes selection. For the databases expansion, besides the SMOTE technique, it was implemented the mega-trend-diffusion (MTD) method. The *M5Rules* algorithm was utilized to find models of estimation of the need of agricultural gypsum. The results showed that the elapsed time after the gypsum application (epoch), the saturation through Ca and the saturation through Mg in the capacity of effective cations exchange (CTCe) of soil were the attributes which presented the narrower correlations with the dose estimation of gypsum. The work identified four possible models for the estimation of the need of agricultural gypsum, showing that the *M5Rules* algorithm was efficient for such prediction. The MTD method presented positive results because increased the correlation coefficient and reduced the average absolute error.

**Keywords:** gypsum, data mining, no-till cropping system, feature selection, mega-trend-diffusion

## **LISTA DE FIGURAS**

|                                                       |    |
|-------------------------------------------------------|----|
| Figura 1 – Processo de KDD .....                      | 23 |
| Figura 2 – MD como campo interdisciplinar.....        | 24 |
| Figura 3 – Definição da tarefa de classificação ..... | 26 |

## LISTA DE TABELA

|                                                                                                                 |    |
|-----------------------------------------------------------------------------------------------------------------|----|
| Tabela 1 – Recomendação de gesso baseado no teor de argila segundo recomendação do Estado de Minas Gerais ..... | 20 |
| Tabela 2 – Análises químicas do solo para as três áreas de estudo .....                                         | 43 |
| Tabela 3 – Conjunto de dados de treinamento selecionadas aleatoriamente.....                                    | 49 |
| Tabela 4 – Representação de um registro para alguns atributos presentes na base de dados .....                  | 50 |
| Tabela 5 – Experimentos da primeira fase .....                                                                  | 50 |
| Tabela 6 – Experimentos da segunda fase.....                                                                    | 51 |
| Tabela 7 – Experimentos da segunda fase.....                                                                    | 52 |
| Tabela 8 – Resultados do $r$ e EMA na base BD1 com subdivisão C e D.....                                        | 58 |
| Tabela 9 – Resultados do $r$ e EMA na base BD2 com subdivisão C e D.....                                        | 59 |
| Tabela 10 – Resultados do $r$ e EMA na base BD3 com subdivisão C e D.....                                       | 59 |
| Tabela 11 – Resultados do $r$ e EMA na base BD3 com subdivisão E e F.....                                       | 62 |
| Tabela 12 – Comparativo das subdivisões G e H para as bases BD1, BD2 e BD3.....                                 | 65 |
| Tabela 13 – Aplicação do método MTD na base BD1 com subdivisão G.....                                           | 76 |
| Tabela 14 – Aplicação do método MTD na base BD2 com subdivisão G.....                                           | 77 |
| Tabela 15 – Aplicação do método MTD na base BD3 com subdivisão G.....                                           | 77 |
| Tabela 16 – Aplicação do método MTD na base BD3 com subdivisão C .....                                          | 78 |
| Tabela 17 – Aplicação do método MTD na base BD3 com subdivisão F.....                                           | 78 |

## LISTA DE QUADROS

|                                                                                   |    |
|-----------------------------------------------------------------------------------|----|
| Quadro 1 – Estimativa para aplicação de gesso utilizando a Regra (2) .....        | 55 |
| Quadro 2 – Estimativa para aplicação de gesso utilizando a Regra (4) .....        | 57 |
| Quadro 3 – Estimativa para aplicação de gesso utilizando a Regra (6) .....        | 60 |
| Quadro 4 – Estimativa para aplicação de gesso utilizando a Regra (7) .....        | 61 |
| Quadro 5 – Estimativa para aplicação de gesso utilizando a Regra (8) .....        | 63 |
| Quadro 6 – Estimativa para aplicação de gesso utilizando a Regra (10) .....       | 64 |
| Quadro 7 – Estimativa para aplicação de gesso utilizando a Regra (13) .....       | 66 |
| Quadro 8 – Estimativa para aplicação de gesso utilizando a Regra (16) .....       | 68 |
| Quadro 9 – Estimativa para aplicação de gesso utilizando a Regra (18) .....       | 69 |
| Quadro 10 – Características das melhores regras geradas.....                      | 72 |
| Quadro 11 – Nova estimativa para aplicação de gesso utilizando a Regra (4) .....  | 72 |
| Quadro 12 – Nova estimativa para aplicação de gesso utilizando a Regra (6) .....  | 73 |
| Quadro 13 – Nova estimativa para aplicação de gesso utilizando a Regra (10) ..... | 73 |
| Quadro 14 – Nova estimativa para aplicação de gesso utilizando a Regra (16) ..... | 74 |

## LISTA DE SIGLAS

|       |                                                   |
|-------|---------------------------------------------------|
| SPD   | Sistema Plantio Direto                            |
| CTCe  | Capacidade de Troca Catiônica efetiva             |
| MD    | Mineração de Dados                                |
| KDD   | <i>Knowledge Discovery in Databases</i>           |
| MTD   | <i>Mega-trend-diffusion</i>                       |
| ACP   | Análise de Componentes Principais                 |
| ACPS  | Análise de Componentes Principais Supervisionado  |
| SMOTE | <i>Synthetic Minority Over-Sampling Technique</i> |
| EMA   | Erro Médio Absoluto                               |
| RNA   | Rede Neural Artificial                            |

## SUMÁRIO

|           |                                                                       |           |
|-----------|-----------------------------------------------------------------------|-----------|
| <b>1.</b> | <b>INTRODUÇÃO .....</b>                                               | <b>13</b> |
| 1.1.      | Objetivos .....                                                       | 14        |
| 1.1.1.    | Objetivos Específicos .....                                           | 14        |
| <b>2.</b> | <b>REVISÃO DA LITERATURA .....</b>                                    | <b>16</b> |
| 2.1.      | Acidez do Solo e sua Correção nos Sistemas de Produção Agrícola.....  | 16        |
| 2.2.      | Gesso Agrícola .....                                                  | 17        |
| 2.2.1.    | Efeitos do gesso agrícola no solo .....                               | 17        |
| 2.2.2.    | Recomendações de gesso agrícola .....                                 | 19        |
| 2.2.3.    | Comportamento das culturas frente à aplicação de gesso agrícola ..... | 20        |
| 2.3.      | Mineração de Dados e Descoberta de Conhecimento em Base de Dados ...  | 22        |
| 2.4.      | Tarefas de MD.....                                                    | 25        |
| 2.4.1.    | Regressão .....                                                       | 27        |
| 2.5.      | Mineração de Dados na Agricultura.....                                | 29        |
| 2.6.      | Mineração de Dados em Pequenas Bases .....                            | 32        |
| 2.7.      | Pré-processamento .....                                               | 35        |
| 2.7.1.    | Seleção de atributos.....                                             | 35        |
| 2.7.1.1.  | Análise de componentes principais (ACP) .....                         | 36        |
| 2.7.2.    | Transformação das bases de dados.....                                 | 39        |
| 2.7.2.1.  | SMOTE .....                                                           | 39        |
| 2.7.2.2.  | Método mega-trend-diffusion .....                                     | 39        |
| 2.8.      | Pós-processamento .....                                               | 41        |
| 2.8.1.    | Interpretação dos resultados .....                                    | 41        |
| <b>3.</b> | <b>METODOLOGIA .....</b>                                              | <b>42</b> |
| 3.1.      | Áreas de estudo .....                                                 | 42        |
| 3.2.      | Sobre os atributos .....                                              | 43        |
| 3.3.      | Organização das bases de dados.....                                   | 44        |
| 3.4.      | Subdivisão das bases de dados .....                                   | 44        |
| 3.5.      | Etapa de pré-processamento.....                                       | 45        |
| 3.5.1.    | Seleção de atributos.....                                             | 45        |
| 3.5.2.    | Transformação das bases de dados.....                                 | 47        |
| 3.5.2.1.  | SMOTE .....                                                           | 47        |
| 3.5.2.2.  | Método mega-trend-diffusion .....                                     | 47        |

|            |                                                                 |           |
|------------|-----------------------------------------------------------------|-----------|
| 3.5.2.2.1. | Exemplo numérico .....                                          | 48        |
| 3.6.       | Etapa de processamento (MD) .....                               | 49        |
| 3.7.       | Etapa de pós-processamento .....                                | 49        |
| 3.8.       | Experimentos realizados .....                                   | 50        |
| 3.8.1.     | Primeira fase .....                                             | 50        |
| 3.8.2.     | Segunda fase .....                                              | 50        |
| 3.8.2.     | Terceira fase .....                                             | 51        |
| <b>4.</b>  | <b>RESULTADOS E DISCUSSÃO .....</b>                             | <b>53</b> |
| 4.1.       | Estimativa de gesso agrícola .....                              | 53        |
| 4.1.1.     | Primeira fase.....                                              | 53        |
| 4.1.2.     | Segunda fase .....                                              | 54        |
| 4.1.3      | Terceira fase .....                                             | 66        |
| 4.1.3.1    | Seleção de atributos com matriz de covariância S .....          | 66        |
| 4.1.3.2    | Seleção de atributos com matriz de covariância $\Omega''$ ..... | 68        |
| 4.2        | Aplicação das melhores regras em um conjunto de testes .....    | 71        |
| 4.3        | Ampliação da base de dados com o método MTD.....                | 75        |
| <b>5.</b>  | <b>CONCLUSÕES .....</b>                                         | <b>81</b> |
| <b>6.</b>  | <b>TRABALHOS FUTUROS.....</b>                                   | <b>82</b> |
| <b>7.</b>  | <b>REFERÊNCIAS BIBLIOGRÁFICAS .....</b>                         | <b>83</b> |

## 1. INTRODUÇÃO

A acidez do subsolo é nociva ao crescimento de raízes das plantas e afeta a produtividade das culturas. A calagem é a prática mais comumente empregada para resolver os problemas da acidez por aumentar o pH do solo, eliminar os efeitos tóxicos causados pelo alumínio (Al) e manganês (Mn), elevar os teores de cálcio (Ca) e magnésio (Mg), e aumentar a disponibilidade de grande parte dos nutrientes no solo. A calagem profunda poderia ser uma solução para a melhoria das condições de acidez do subsolo. Entretanto, a incorporação profunda de calcário no solo é uma prática onerosa, demanda alta potência das máquinas e intenso revolvimento, sendo uma alternativa de pouco interesse, especialmente para solos manejados no sistema plantio direto (SPD).

O SPD com rotação diversificada de culturas tem se destacado como uma das alternativas mais eficazes para a produção agrícola de regiões tropicais e subtropicais por reduzir perdas de solo e de nutrientes por erosão. O SPD tem apresentado rápido crescimento em área cultivada no Brasil, apresentando atualmente mais de 25 milhões de ha em tal sistema de manejo. No Estado do Paraná, cerca de 90% da área de produção de grãos das culturas é manejada no SPD. Em solos sob SPD, a correção da acidez do solo é feita por meio da aplicação de calcário na superfície, sem incorporação. Devido ao fato de a reação do calcário geralmente ser limitada ao local de sua aplicação, a calagem na superfície em SPD apresenta pouca ação na correção da acidez do subsolo.

O gesso agrícola ( $\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$ ), um subproduto da indústria de fertilizantes fosfatados, é uma excelente fonte de Ca e enxofre (S), e pode atenuar o efeito tóxico do Al em camadas do subsolo. Embora não seja um corretivo da acidez do solo, a aplicação de gesso agrícola na superfície seguida por lixiviação para subsolos ácidos resulta em melhor desenvolvimento radicular e maior absorção de água e nutrientes pelas raízes das plantas (RITCHEY *et al.*, 1980; CAIRES *et al.*, 2004; SORATTO e CRUSCIOL, 2008a) por aumentar a concentração de Ca e reduzir a atividade do Al em camadas do subsolo.

Uma das grandes dificuldades ainda existentes com relação ao emprego de gesso na agricultura é dimensionar adequadamente as doses a serem aplicadas em cada situação, com base na análise química do solo, visando alcançar os efeitos desejados na melhoria das condições químicas do solo e no aumento da produtividade das culturas. Tal questão motiva a realização de estudos envolvendo o uso do gesso agrícola, e a análise de bases de dados contendo registros referentes à aplicação do gesso agrícola e

também de atributos químicos do solo e de produtividade das culturas ao longo do tempo, pode ser uma forma de aperfeiçoar as recomendações da necessidade de gesso.

Uma abordagem computacional que pode ser utilizada para a análise de bases de dados é a técnica de Aprendizado de Máquina, denominada de Mineração de Dados (MD).

A MD permite a realização da análise de uma base de dados, buscando padrões e correlações existentes na mesma, preferencialmente com característica de novidade. Além disso, esses padrões e correlações devem possuir um grau de certeza e confiabilidade elevadas, para isso é importante a interação com o especialista humano na área de domínio. De modo que seja possível analisar os resultados obtidos tanto do ponto de vista numérico quanto do ponto de vista da área de interesse.

Dentre os conceitos de MD, a qual consiste basicamente em buscar conhecimento dentro de uma grande base de dados, destaca-se a necessidade de uma extensa base de dados. Na área agrícola, dificilmente encontram-se bases de dados extensas, como é o caso de bases de dados experimentais envolvendo aplicações de gesso. Entretanto, há estudos abordando novas técnicas de MD focadas para pequenos conjuntos de dados.

Esse trabalho foi realizado com a motivação de estudar e explorar essa nova área da MD focada em pequenas bases, visando contribuir para o aprimoramento das recomendações de gesso na agricultura e buscando melhor compreensão dos efeitos de sua aplicação.

## **1.1. Objetivos**

O objetivo geral desse trabalho foi realizar a MD em uma base referente à aplicação de gesso agrícola em três áreas distintas sob SPD, buscando correlações entre os atributos químicos do solo e a estimativa da necessidade de gesso.

### **1.1.1. Objetivos Específicos**

- Obter melhor compreensão dos efeitos da aplicação de gesso agrícola nos atributos químicos do solo utilizando o processo de Descoberta de Conhecimento em Bases de Dados (DCBD ou KDD – do inglês *Knowledge Discovery in Databases* ).

- Identificar os atributos químicos do solo que apresentam correlações mais estreitas com a necessidade de gesso agrícola usando técnicas de seleção de atributos na etapa de pré-processamento.
- Definir modelos para a estimativa da dose gesso agrícola a ser aplicada.
- Explorar técnicas de MD em pequenos conjuntos, adequando-as ao domínio do problema estudado.

## 2. REVISÃO DA LITERATURA

### 2.1. Acidez do Solo e sua Correção nos Sistemas de Produção Agrícola

No Brasil, a maior parte dos solos agrícolas apresenta acidez excessiva (LOPES *et al.*, 2004). A acidificação do solo é um processo natural resultante da lixiviação de cátions básicos solúveis (Ca, Mg e potássio(K)) seguida pela sua substituição por cátions ácidos (H e Al) no complexo de troca catiônica (ZIGLIO *et al.*, 1999). Os solos podem ser ácidos naturalmente devido à pobreza do material de origem em cátions básicos (MARQUES, 2008). Além disso, a acidez do solo pode ser aumentada devido a processos de formação que favorecem a remoção e a perda dos cátions básicos. De qualquer forma, a acidificação se inicia ou se acentua com a remoção de bases da superfície dos colóides do solo (LOPES *et al.*, 1991).

Os solos ácidos contêm elevados teores de Al e Mn trocáveis, baixa saturação por bases e baixo valor de pH (VELOSO *et al.*, 1992). Por isso, a acidez do solo é um dos principais fatores que limita a produtividade agrícola. A acidez do solo também proporciona restrição ao crescimento radicular e à absorção de água e nutrientes pelas culturas (CAIRES *et al.*, 1998). A deficiência do Ca e a toxicidade por Al têm sido as principais barreiras químicas para o crescimento radicular das plantas em subsolos ácidos (RIOS e PEARSON 1964, ADAMS e LUND 1966 apud QUAGGIO *et al.*, 1993).

Para que o sistema radicular explore maior volume de solo e, assim, as plantas possam absorver água e nutrientes necessários para o seu desenvolvimento, é importante corrigir todo o perfil do solo (RAMOS *et al.*, 2006). O material mais utilizado como corretivo da acidez do solo é o calcário, pois além de neutralizar os solos ácidos é capaz de fornecer nutrientes, principalmente Ca e Mg, para as plantas. Para maior eficiência na correção da acidez, o calcário deve ser incorporado ao solo (RAMOS *et al.*, 2006 ; SORATTO e CRUSCIOL, 2008a). Porém, a calagem profunda com incorporação encontra problemas do ponto de vista técnico e, até mesmo, econômico (NEIS, 2009), pois exige máquinas e equipamentos caros, tornando a prática onerosa (CAIRES *et al.*, 1998). Mesmo assim, essa técnica tem demonstrado bons resultados em sistema convencional de preparo do solo (MASCARENHAS *et al.*, 1969, 1976; RAIJ *et al.*, 1983; QUAGGIO *et al.*, 1993).

No Brasil, o SPD teve um rápido crescimento em área cultivada a partir da década de 1990 (CAIRES *et al.*, 1999). No Estado do Paraná, os sistemas de produção já começaram a mudar a partir da década de 1970 (MUZILLI, 1981 apud CAIRES *et*

*al.*, 2000). O SPD consolidou-se como uma técnica conservacionista, sendo amplamente utilizada no Brasil. Estima-se que o SPD ocupe mais de 25 milhões de hectares no Brasil, ou seja, cerca de 50% da área com culturas anuais no país (CRUZ *et al.*, 2010).

A calagem com incorporação não pode ser utilizada no SPD, pois este é um sistema conservacionista que, por preconizar o não revolvimento do solo, promove acúmulo de matéria orgânica, principalmente em sua superfície, o que propicia maior disponibilidade de água e nutrientes para as culturas. Desse modo, pesquisas vêm sendo realizadas para auxiliar na compreensão do comportamento da calagem (CAIRES *et al.*, 1998; CAIRES *et al.*, 2000; SORATTO e CRUSCIOL, 2008b), uma vez que, em áreas sob SPD, o calcário é aplicado na superfície do solo sem incorporação.

Como os corretivos de acidez são pouco solúveis em água, a reação do calcário aplicado na superfície em SPD é mais lenta e restrita às camadas superficiais do solo (CAIRES *et al.*, 1999, 2004; KAMINSKI *et al.*, 2005; CAIRES *et al.*, 2006). Com isso, o crescimento radicular e a nutrição das plantas podem ser comprometidos, caso a acidez do subsolo não seja adequadamente corrigida (CAIRES *et al.*, 2004).

## **2.2. Gesso Agrícola**

O gesso agrícola ( $\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$ ) é um subproduto das indústrias de fertilizantes fosfatados que tem sido empregado na agricultura como fonte de Ca e S para as plantas, e também como condicionador de solo (RAMPIM, 2008). O gesso agrícola não é um corretivo da acidez do solo, ou seja, a sua utilização não aumenta o pH do solo. Desse modo, o gesso não substitui o calcário, porém pode complementar seu efeito reduzindo a fitotoxicidade do Al em profundidade (VELOSO *et al.*, 1992).

### **2.2.1. Efeitos do gesso agrícola no solo**

O gesso agrícola, quando aplicado na superfície do solo, movimenta-se ao longo do perfil sob influência do excesso de umidade (CAIRES *et al.*, 1999). O aumento de Ca em profundidade reduz a saturação e a atividade do Al (DIAS, 1992; FARIA *et al.*, 2003), favorecendo o enraizamento profundo das culturas em camadas do subsolo. Com isso, o gesso atua sobre os dois fatores que afetam o crescimento radicular das plantas – a deficiência de Ca e a toxicidade por Al (RAIJ *et al.*, 1998), e não exige revolvimento do solo. Como o calcário, quando aplicado na superfície do solo, tem menor influência na correção da acidez do subsolo, o gesso pode ser um insumo complementar

importante à calagem superficial ao melhorar o ambiente radicular de camadas mais profundas do solo (RAIJ, 2008).

Após o trabalho pioneiro de Ritchey *et al.* (1980) que demonstrou efeito do gesso na melhoria das condições de acidez do subsolo, permitindo maior aprofundamento das raízes do milho, o interesse por estudos envolvendo a aplicação de gesso agrícola intensificaram-se no Brasil. A partir daí, o uso de gesso agrícola se tornou polêmico em decorrência das frequentes contradições sobre suas propriedades confundindo-as com as propriedades dos corretivos da acidez do solo (BORKERT *et al.*, 1987; ALCARDE, 1988). Essas confusões foram originadas da relação existente entre acidez e Al, já que o gesso é capaz de agir sobre a toxicidade por Al (CAIRES *et al.*, 1998, 1999; SORATTO e CRUSCIOL, 2008a; NEIS *et al.*, 2010). Acontece, porém, que mesmo diminuindo a toxicidade por Al, o gesso praticamente não influencia o pH do solo (CAIRES *et al.*, 1998; QUAGGIO *et al.*, 1993; RAIJ *et al.*, 1998; VELOSO *et al.*, 1992). A aplicação de gesso pode apenas ocasionar pequena elevação no pH em camadas do subsolo (CAIRES *et al.*, 1999, 2003, 2004), devido a uma reação de troca de ligantes na superfície das partículas do solo, envolvendo óxidos hidratados de ferro (Fe) e Al, com o  $\text{SO}_4^{2-}$  deslocando  $\text{OH}^-$  e, desse modo, promovendo neutralização parcial da acidez (REEVE e SUMNER, 1972 apud CAIRES *et al.*, 2004). Ernani *et al.* (2001) observaram que a adição de gesso agrícola diminuiu o pH do solo e da solução percolada, não afetou o Al trocável e percolou mais Ca, Mg e Al que os corretivos de acidez.

Um problema frequentemente encontrado com a aplicação de gesso em solos é a intensa movimentação de bases trocáveis que a adição de gesso proporciona. Além do Ca, a aplicação de gesso também pode movimentar Mg e K das camadas superficiais para as do subsolo. Caires *et al.* (1998) verificaram aumento do teor de Ca em todo perfil do solo e lixiviação de Mg das camadas superficiais para as camadas mais profundas, após oito meses de aplicação do gesso em um Latossolo Vermelho textura média sob SPD. Caires *et al.* (2003) também observaram redução de Mg trocável nas camadas superficiais do solo e aumento no teor desse nutriente nas camadas do subsolo, após oito meses de aplicação de gesso em um Latossolo Vermelho textura argilosa sob SPD. Em outro trabalho, Quaggio *et al.* (1993) verificaram substituição de Mg por Ca nas camadas superficiais do solo com a aplicação de gesso.

Dependendo da quantidade de gesso aplicado e do tempo decorrido após a aplicação, a lixiviação do Mg pode ser ainda mais acentuada (CAIRES *et al.*, 1998,

1999). Assim, quando o gesso é aplicado em doses elevadas no solo devem ser desenvolvidas estratégias para minimizar as perdas de Mg trocável. Dessa forma, recomenda-se utilizar o gesso combinado com o uso de calcário dolomítico, e não de forma isolada, de modo a garantir maiores teores de Mg nas camadas superficiais do solo (CAIRES *et al.*, 1999, 2003). No caso do K, a lixiviação provocada pelo uso de gesso também pode ocorrer, dependendo do tipo de solo (SOUSA e RITCHEY, 1986), mas tem sido relativamente pequena em SPD (CAIRES *et al.*, 1998, 2002, 2004, 2011). É importante considerar que tanto a redução dos teores de Mg, como a redução dos teores de K, nem sempre são consideradas perdas (RAIJ *et al.*, 1998). Somente com a lixiviação de sulfato para fora do alcance das raízes poderão ocorrer perdas dessas bases.

### 2.2.2. Recomendações de gesso agrícola

Segundo Raij (2008) as recomendações da necessidade de gesso existentes são consideradas conservadoras, pois leva em consideração não apenas a preocupação em não impor gastos elevados aos produtores, mas também a preocupação em evitar perdas de Mg e K por lixiviação. Por outro lado, tais recomendações não contemplam maior aprofundamento do efeito do gesso, o que gera dúvidas sobre o impacto efetivo de sua aplicação, ao longo do perfil do solo.

Em solos da região dos cerrados (Sousa *et al.*, 2005), a recomendação da necessidade da dose de gesso é dada pelo teor de argila do solo, sendo a Equação (1) para culturas anuais, e a Equação (2) para culturas perenes.

$$N.G^1(\text{kg/ha}) = 50 \times \text{Argila} (\%) \quad (1)$$

$$N.G(\text{kg/ha}) = 75 \times \text{Argila} (\%) \quad (2)$$

O teor de argila é um atributo do solo que pode ser utilizado para a quantificação da gessagem, porém sua consideração não dá sustentação para o uso em solos de propriedades eletroquímicas diferentes, com capacidades diversas de retenção de sulfato (Raij, 2008).

Sousa *et al.* (2005) afirmam que também é possível determinar as doses de gesso agrícola com base na classificação textural do solo, sendo para culturas anuais, as doses de 700, 1200, 2200 e 3200 (kg/ha), e quanto as culturas perenes, as doses são de 1050,

---

<sup>1</sup> N.G = Necessidade de gesso

1800, 3000 e 4800 (kg/ha) para as texturas arenosa, média, argilosa e muito argilosa, respectivamente.

No Estado de Minas Gerais (Ribeiro *et al.*, 1999) a recomendação da dose de gesso agrícola é dada quando na camada superficial (20 a 40 cm ou 30 a 60 cm) houver  $\text{Ca}^{2+}$  igual ou inferior a  $4 \text{ mmol/dm}^3$  e/ou  $\text{Al}^{3+}$  maior do que  $5 \text{ mmol/dm}^3$  e/ou, ainda, quando a saturação por Al é maior do que 30%. Nesse caso a necessidade da dose de gesso, para uma camada subsuperficial de 20 cm de espessura, é dada pela Tabela 1.

Tabela 1 – Recomendação de gesso baseado no teor de argila segundo recomendação do Estado de Minas Gerais

| <b>Argila (g/kg)</b> | <b>N.G (t/ha)</b> |
|----------------------|-------------------|
| 0 a 150              | 0,0 a 0,4         |
| 150 a 350            | 0,4 a 0,8         |
| 350 a 600            | 0,8 a 1,2         |
| 600 a 1000           | 1,2 a 1,6         |

Raij (2008) afirma ainda que as quantidades recomendadas não são elevadas e poderiam ser aumentadas em solos com capacidade considerável de retenção de sulfato. Como o gesso pode apresentar um prolongado efeito residual, de muitos anos, existe um resultado econômico favorável podendo se acentuar com o passar do tempo. O autor comenta que apesar da existência de tais recomendações, ainda faltam informações sobre a influência de diversos atributos do solo para uma recomendação de gesso mais “científica”, destacando-se adsorção de sulfato, profundidade a ser atingida, valores críticos de saturação por Al, etc.

### 2.2.3. Comportamento das culturas frente à aplicação de gesso agrícola

A aplicação de gesso agrícola aumenta os teores de Ca nas folhas das culturas trigo (CAIRES *et al.*, 1999, 2003, 2004). O trigo apresenta resposta controversa quanto à influência do gesso na produtividade de grãos, pois há trabalhos indicando ausência de resposta e outros indicando aumento na produtividade da cultura com a adição de gesso. Por exemplo, Caires *et al.* (2002) obtiveram acréscimos na produtividade de grãos de trigo em resposta à aplicação de gesso em SPD, a qual também aumentou a concentração foliar de Ca e S. Outro trabalho em que houve resposta positiva na produtividade de grãos de trigo com o uso de gesso em SPD foi o de Rampim (2008).

Caires *et al.* (1999) avaliaram a produção das culturas do milho, soja e trigo em função das alterações químicas do solo pela aplicação de calcário e gesso na superfície em SPD e constataram que o uso de gesso ocasionou aumento na produtividade de

grãos somente da cultura do milho. Nesse caso, o incremento na produtividade de milho com a aplicação de gesso esteve relacionado com o fornecimento de S, o aumento de Ca trocável em todo o perfil do solo, a redução do Al trocável e o aumento da relação Ca/Mg do solo. Em outro estudo realizado com o objetivo de estudar as alterações químicas do solo e a resposta do milho à calagem e aplicação de gesso em SPD verificou-se que a combinação da calagem com o uso de gesso ocasionou aumento da ordem de 17% na produtividade de grãos de milho (CAIRES *et al.*, 2004). Em um estudo mais recente, Caires *et al.* (2011) também encontraram aumento na produtividade de grãos do milho após aplicação de gesso em SPD. Em sistema convencional de preparo do solo, Raij *et al.* (1998) avaliaram o efeito do calcário e do gesso na produção de cultivares de milho tolerante e sensível ao Al e verificaram que, independentemente da aplicação de calcário, o gesso favoreceu a produtividade do cultivar sensível ao Al. Portanto, a cultura do milho tem apresentado respostas positivas à aplicação de gesso em diversas condições de solo, clima e sistema de cultivo.

Para a cultura da cevada, Caires *et al.* (2001) verificaram que a aplicação de gesso em SPD aumentou a absorção de N, P, K, Ca e S pelas plantas em condições de deficiência de água e proporcionou acréscimos na produtividade de grãos. Isso ocorreu devido ao aumento nos teores de Ca trocável, da relação de Ca/Mg e do  $\text{S-SO}_4^{2-}$  disponível no solo. Destaca-se que, nesse estudo, houve efeito positivo do gesso no fornecimento de S sobre a produtividade de grãos de cevada, em solo com  $10,4 \text{ mg dm}^{-3}$  de  $\text{S-SO}_4^{2-}$  na camada de 0–20 cm. Como o crescimento radicular da cevada não foi alterado com a adição de gesso, não foram encontradas relações significativas entre os atributos de crescimento de raízes e os atributos químicos do solo relacionados com a acidez.

No caso da soja, Nogueira e Melo (2003) destacaram a importância do S como nutriente para o aumento da cultura. Desse modo, a aplicação do gesso agrícola pode aumentar a produtividade da soja pelo fornecimento de S (MASCARENHAS *et al.*, 1986 apud NOGUEIRA e MELO, 2003). Porém, no trabalho de Nogueira e Melo (2003), as doses de gesso não influenciaram a produtividade de grãos de soja. A ausência de resposta na produtividade de soja com o uso de gesso evidenciou que os teores de S no solo não foram limitantes para a cultura, provavelmente em decorrência do S mineralizado pela decomposição da matéria orgânica do solo. Concordando com esses resultados, Caires *et al.* (1998) também não constatarem resposta positiva da cultura da soja para o gesso aplicado na superfície do solo em SPD. Inclusive, com a

dose máxima de gesso aplicada nesse estudo ( $12 \text{ t ha}^{-1}$ ), houve redução significativa na produtividade de grãos de soja em relação ao tratamento controle sem gesso em razão da lixiviação de Mg trocável no solo.

Em outro estudo realizado em sistema convencional de preparo do solo, Quaggio *et al.* (1993) também verificaram ausência de resposta da soja à aplicação de gesso. Neis *et al.* (2010) constaram que não houve alteração no rendimento de grãos da soja com a aplicação de gesso, tanto no SPD quanto no sistema convencional de preparo. Segundo Caires *et al.* (2003), a ausência de resposta positiva da soja ao uso de gesso pode estar ligada ao fato de o crescimento radicular da soja, na ausência de déficit hídrico, não ser influenciado pela redução da saturação por Al no subsolo. Além disso, a soja parece apresentar maior exigência em Mg e menor habilidade no aproveitamento desse nutriente em camadas do subsolo, quando comparada ao trigo e ao milho.

Mesmo havendo informações controvertidas referentes ao uso do gesso agrícola e respostas diferenciadas das culturas com a sua aplicação, o gesso agrícola é considerado um insumo capaz de melhorar o ambiente radicular de solos ácidos e beneficiar o crescimento radicular e a produtividade das culturas. Entretanto, ainda há dúvidas sobre quais as situações em que se podem esperar efeitos favoráveis da aplicação de gesso na produção das culturas e quanto ao método de recomendação do produto (CAIRES *et al.*, 2011). Assim, é de grande importância o estudo envolvendo a análise de base de dados contendo registros referentes à aplicação de gesso agrícola, de modo a auxiliar na compreensão de seu comportamento no solo visando a definição de doses adequadas para a sua recomendação.

### **2.3. Mineração de Dados e Descoberta de Conhecimento em Base de Dados**

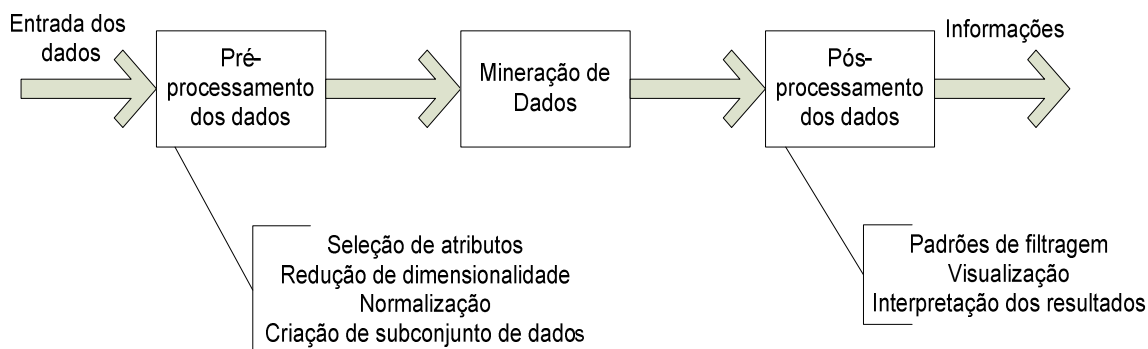
A MD é responsável por encontrar padrões relevantes entre os dados analisados e, dessa forma, extrair algum conhecimento implícito nos mesmos que pode ser aplicado em diversas áreas de estudo. Segundo Tan *et al.* (2009), o conceito de MD consiste no processo de descoberta automática de informações úteis em grandes depósitos de dados. Conceitos similares a esse podem ser encontrados na literatura, como em Dias (2002) que explica que as técnicas de MD são aplicadas em sistemas de descoberta de conhecimento em grandes bancos de dados com objetivo de extrair informações estratégicas.

A MD é a etapa mais importante de um processo denominado Descoberta de Conhecimento em Bancos de Dados, ou *Knowledge Discovery in Databases* (KDD).

Esse processo leva a conversão de dados brutos em informações úteis que consiste de uma série de etapas, desde a transformação, o pré-processamento, a mineração em si, até o pós-processamento dos resultados obtidos (TAN *et al.*, 2009).

O KDD tem como objetivo a extração de conhecimento de alto nível a partir de dados de baixo nível no contexto de grandes conjuntos de dados (FAYYAD *et al.*, 1996). Bürkle (2006) explica que o KDD é um processo iterativo, pois pode ser repetido mais de uma vez com o objetivo de obter resultados melhores na próxima iteração, sendo considerado também um processo interativo, pois envolve a participação de diferentes classes de usuários, como (i) os especialistas do domínio, os quais são responsáveis por fornecer apoio para a execução do processo, (ii) os analistas, os quais são responsáveis pela execução do processo e (iii) o usuário final que irá utilizar o conhecimento extraído do processo. A Figura 1 representa uma visão geral sobre o processo de KDD.

Figura 1 – Processo de KDD, segundo Tan *et al.* (2009)



Para que seja possível transformar os dados de forma adequada para a execução da MD é essencial, primeiramente, obter a compreensão sobre o domínio de aplicação e dos objetivos da análise a ser realizada. Esta compreensão é de grande importância, pois uma das grandes dificuldades do processo de KDD é identificar qual tarefa de MD é mais indicada em função dos tipos de dados e da meta a ser atingida (GUIMARÃES, 2005).

Bases de dados composta por dados reais podem conter dados incompletos e ruidosos, no entanto, a partir da compreensão sobre o domínio de aplicação é possível aplicar algumas técnicas da etapa de pré-processamento que visam eliminar essas inconsistências (BÜRKLE, 2006). Ainda na etapa de pré-processamento objetiva-se a transformação dos dados de entrada brutos (baixo nível) em um formato adequado para as análises subsequentes (TAN *et al.*, 2009).

Dentre as ações que podem ser tomadas fase de pré-processamento do KDD está as técnicas de seleção de atributos, limpeza da base quanto à detecção e remoção de dados ruidosos e/ou duplicados, transformação dos dados e seleção de registros que sejam relevantes à tarefa de MD (KANTARDZIC, 2003; TAN *et al.*, 2009).

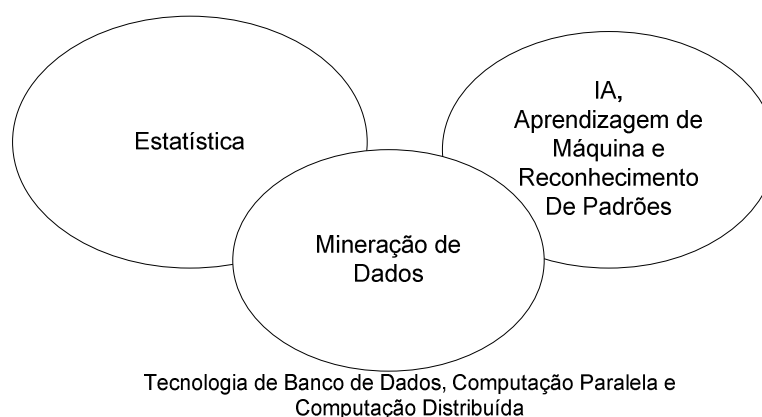
Posterior a etapa de pré-processamento, há a etapa de MD, que é o processamento propriamente dito, esta etapa consiste na aplicação de algoritmos específicos para extrair padrões de dados. A etapa de MD é responsável por definir qual tarefa de MD será aplicada e posteriormente qual algoritmo será usado para encontrar os padrões nos dados (GUIMARÃES, 2005).

A aplicação de algum algoritmo de MD sem que exista o tratamento dos dados na etapa de pré-processamento é uma atividade perigosa que pode levar a descoberta de padrões sem sentido e inválidos (FAYYAD *et al.*, 1996).

Bürkle (2006) afirma que os padrões devem ser válidos, ou seja, devem possuir um grau de certeza de modo que os exemplos cobertos sejam aceitáveis. Galvão e Marin (2009), afirmam que, além de novo, útil e válido, os padrões devem ser, eficientes (acurados), genéricos (aplicáveis a vários tipos de dados) e flexíveis (facilmente modificáveis). Assim, somente os resultados com tais características são analisados, interpretados e aplicados ao domínio de estudo.

A MD é um campo interdisciplinar que envolve principalmente as áreas de Estatística, Inteligência Artificial e Aprendizado de Máquina (TAN *et al.*, 2009). A Figura 2 representa a MD como uma confluência de muitas disciplinas.

Figura 2 – MD como campo interdisciplinar, segundo Tan et al. (2009)



Guimarães (2005) afirma que a interdisciplinaridade da MD é essencial, pois as áreas associadas precisam ser aplicadas de forma integrada para a obtenção de bons resultados. A Estatística Clássica, que envolve os conceitos de distribuição normal,

variância, análise de regressão, análise de conjuntos, análise de discriminantes, desvio simples e intervalos de confiança, todos usados para estudar dados e relacionamentos entre eles, é a mais antiga linhagem de descendência da MD, isso porque essas técnicas da Estatística Clássica são a base para a maioria das tecnologias de MD.

Quanto à área de Inteligência Artificial, Guimarães (2005) afirma que esta área é construída a partir dos fundamentos da heurística, em oposto a estatística, de modo que tenta imitar a forma como o ser humano pensa na resolução de problemas estatísticos. Por fim, quanto à área de Aprendizagem de Máquina, a autora afirma que esta é a terceira e última linhagem da MD que pode ser descrita como uma associação entre a Estatística e a Inteligência Artificial, pois combina a heurística e a análise estatística.

Guimarães (2005) afirma ainda que a MD consiste em mais do que uma simples consulta em um banco de dados, no sentido em que permite a exploração e inferência de informação útil a partir dos dados descobrindo relacionamentos implícitos entre os dados. Para isso, na MD são definidas as tarefas e algoritmos que serão utilizados.

Finalmente, na etapa de pós-processamento é realizada a interpretação dos padrões encontrados, buscando analisar se o resultado obtido é útil (capaz de trazer benefício ao usuário) e com caráter de novidade (desconhecido pelo usuário).

## **2.4. Tarefas de MD**

As tarefas de MD podem ser agrupadas em modelos preditivos, que serão julgados pela precisão da previsão empírica sobre um conjunto de teste, e modelos descritivos que podem ser avaliados ao longo das dimensões de precisão de previsão, novidade, utilidade e compreensibilidade do modelo ajustado (FAYYAD *et al.*, 1996).

Segundo Pimentel e Omar (2006), as tarefas de MD correspondem a classes de problemas que são determinadas de acordo com o tipo de conhecimento a ser minerado e também dos objetivos desejáveis para a solução, e a escolha da técnica e do algoritmo a ser utilizado depende da tarefa que se quer executar. As principais tarefas são a associação, agrupamento, classificação e regressão.

A tarefa de associação consiste em identificar e descrever associações entre atributos no mesmo item ou associações entre itens diferentes que ocorram simultaneamente (GALVÃO e MARIN, 2009), ou seja, esta tarefa permite identificar as relações presentes entre diferentes atributos em um registro (GUIMARÃES, 2005).

Também se pode encontrar na literatura o termo “Regras de associação” como denominação para a tarefa de associação (CASTANHEIRA, 2008), isso ocorre, pois os

relacionamentos descobertos podem ser representados na forma de regras de produção do tipo:

$$\text{SE } (X) \text{ ENTÃO } (Y),$$

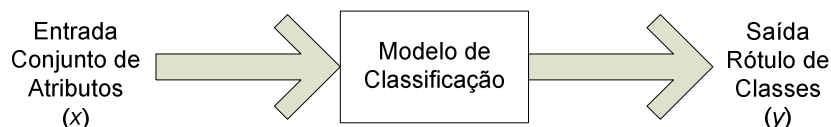
em que  $X$  e  $Y$  são conjuntos de atributos, com  $X \cap Y = \emptyset$  (GUIMARÃES, 2005).

A tarefa de agrupamento, também conhecida como clusterização (do inglês *cluster*), consiste em dividir a base de dados em subconjuntos ou *clusters* (grupos) de dados que compartilhem características em comum. Essas características devem distinguir os elementos em outros *clusters*, tendo como objetivo maximizar a semelhança *intra-cluster* e minimizar a similaridade *inter-cluster* (GALVÃO e MARIN, 2009). Essa tarefa, muitas vezes, é apenas o passo inicial da MD, podendo, posteriormente, ser aplicadas outras tarefas de MD para os *clusters* gerados.

Já a tarefa de classificação mapeia os dados de entrada em um número finito de categorias pré-definidas, também chamadas de classes ou atributo meta, e cada exemplo (registro) pertence a uma classe, de modo que o modelo gerado pelo processo de classificação pode ser aplicado a dados ainda não classificados, a fim de caracterizá-los nas classes definidas (Silva, 2007).

Segundo Tan *et al.* (2009) a definição de classificação é “a tarefa de aprender uma função alvo  $f$  que mapeie cada conjunto de atributos  $x$  para um dos rótulos de classes  $y$  pré-determinados”. Essa função alvo é conhecida informalmente como modelo de classificação. A Figura 3 ilustra a definição acima.

Figura 3 – Definição da tarefa de classificação, segundo Tan et al. (2009)



Esse tipo de tarefa é mais apropriado para prever conjuntos de dados com categorias nominais ou binárias, sendo menos efetivo para categorias ordinais, pois não considera a ordem implícita entre as categorias (TAN *et al.*, 2009). Para Pimentel e Omar (2006), a principal vantagem dessa tarefa é que a forma como o resultado obtido é apresentado é compreensível para a maioria dos humanos. No entanto, não deve eliminar a necessidade do especialista da área de domínio para auxílio na interpretação dos resultados.

Sobre a tarefa de regressão, sendo esta de especial interesse para esse trabalho, será discutida em mais detalhes na seção a seguir.

### 2.4.1. Regressão

A tarefa de regressão é similar à tarefa de classificação. A principal diferença está no atributo a ser predito, que é do tipo contínuo, ao contrário da tarefa de classificação em que o atributo meta é discreto (CASTANHEIRA, 2008). Fayyad *et al.* (1996) afirma que a tarefa de regressão aprende uma função que mapeia um item de dado para um atributo de predição real.

Formalmente, a regressão é o processo para determinar como o atributo  $Y$  é relaciona-se com os atributos  $x_1, x_2, x_3, \dots, x_n$ , em que  $Y$  é o atributo dependente (meta ou saída), e os atributos  $X_i$  são os atributos independentes (entrada) (KANTARDZIC, 2003).

Segundo Castanheira (2008), o modelo deve ter a seguinte forma:

$$Y = f(x_1, x_2, \dots, x_n),$$

em que  $X = \{x_1, x_2, \dots, x_n\}$ , seja o conjunto de atributos de entrada e  $Y$  seja o atributo meta. Então, esse modelo é utilizado para predizer o valor do atributo meta contínuo para um conjunto com  $n$  novas amostras. Nesse caso, pode-se observar os resultados reais e comparar com o que foi previsto.

O objetivo da regressão é determinar o melhor modelo que pode relacionar o atributo meta com os atributos de entrada (preditores) (CASTANHEIRA, 2008). Com disso, é possível predizer o comportamento do atributo meta, bem como identificar as ligações existentes, se houver, entre os atributos de entrada com o atributo de saída (KANTARDZIC, 2003).

Assim como a tarefa de classificação, a regressão também utiliza métodos de indução de regras e Árvores de Decisão. As árvores de regressão são compostas por dois tipos de nós: os nós internos, em que cada nó corresponde a um teste feito em um dos atributos de entrada e os nós folha, em que são feitas as predições do atributo meta. Os nós folha de uma árvore de regressão possuem uma função matemática para predizer o atributo de saída.

Um dos algoritmos empregados para a tarefa de regressão é o *M5Rules*, capaz de gerar regras a partir de modelos de Árvore de Decisão. O algoritmo funciona da seguinte forma: uma árvore de aprendizado é aplicada sobre todo o conjunto de treinamento e uma árvore podada é aprendida. Em seguida, a melhor ramificação (de acordo com alguma heurística) gera uma regra e a árvore é descartada. Todas as instâncias cobertas pela regra são removidas do conjunto de dados, e o processo é aplicado de modo recursivo para os exemplos restantes até que todas as instâncias sejam

cobertas por uma ou mais regras. Ao invés de criar uma única regra de aprendizagem, constrói-se um modelo de árvore completo em cada fase e faz-se da melhor ramificação uma nova regra (HOLMES *et al.*, 1999).

O algoritmo *M5Rules* produz regras de regressão proposicionais no formato SE-ENTÃO, embora não sejam obrigatórias apenas regras contendo condicionais. Seja  $Y$  o atributo meta com valor contínuo,  $X$  o conjunto de atributos de entrada, em que  $X = \{x_1, x_2 \text{ e } x_3\}$ , assumindo ainda que,  $\alpha$ ,  $\beta$ , e  $\gamma$  são os coeficientes de regressão dos atributos e  $\epsilon$  como sendo o valor de ajuste da regra.

$$\text{SE } x_1 > 24,6 \text{ ENTÃO } Y = \alpha * x_1 - \beta * x_2 - \epsilon \quad (\text{a})$$

$$Y = \alpha * x_1 + \beta * x_2 - \gamma * x_3 + \epsilon \quad (\text{b})$$

$$Y = \gamma * x_3 + \epsilon \quad (\text{c})$$

As Regras (a), (b) e (c) são exemplos de regras que poderiam ser geradas utilizando o algoritmo *M5Rules*. Além de a condicional SE não ser obrigatória para a geração de regras com este algoritmo, não é necessário também conter todos os atributos em uma mesma regra, assim como na regra (c), em que se utiliza apenas o atributo  $x_3$ .

O algoritmo *M5Rules* é aplicado em diferentes áreas de estudo, por exemplo, Pina e Pina Filho (2010) utilizaram diferentes técnicas de MD para previsão dos resultados de ensaios de fluência em aços, em função da decomposição química e das condições do ensaio. Apesar do resultado insatisfatório quanto às técnicas utilizadas para o objetivo proposto, o algoritmo *M5Rules* obteve o terceiro menor valor de EMA dentre os quinze algoritmos testados.

Outro exemplo de aplicação do algoritmo *M5Rules* é o projeto de software. Duggal e Singh (2012) utilizaram, dentre outros, o algoritmo *M5Rules* para modelagem de estimativa de esforço de projetos de software. Os resultados demonstraram que o *M5Rules* teve o melhor desempenho e pode ser utilizado para estimativa de esforço de projeto de software.

Pina *et al.* (2011) afirmam que dentre os algoritmos testados, em seu trabalho, o que mostrou-se mais apropriado para o problema de estimação da tenacidade Charpy<sup>2</sup> de soldas metálicas foi o *M5Rules*. Outro trabalho é em Pina *et al.* (2010), cujo objetivo era aplicar técnicas de MD na estimativa da força compressiva do concreto, também

---

<sup>2</sup> O ensaio de choque Charpy é um teste que mede a quantidade de energia absorvida por uma amostra de material, quando submetida à ação de um esforço de choque de valor conhecido (PINA *et al.*, 2011).

testou o algoritmo *M5Rules*. Dentre os algoritmos testados, o *M5Rules* apresentou o segundo menor erro médio quadrático.

Na área agronômica pode-se citar o trabalho de Kotsiantis *et al.* (2007), no domínio da agrometeorologia, que objetivou estimar os valores de temperatura máxima, mínima e média, utilizando, entre outros, o algoritmo *M5Rules*. Na conclusão, os autores afirmaram que os algoritmos de regressão permitem aos especialistas prever os valores mínimo, médio e máximo de temperatura com uma precisão satisfatória.

Alberg *et al.* (2009), propuseram um novo algoritmo de previsão capaz de induzir um modelo para estimativa dado que o valor de saída é contínuo em um conjunto de dados temporais, e para isso utilizaram, entre outros, o algoritmo *M5Rules* no estudo de modo que pudesse ser realizado uma análise comparativa.

## **2.5. Mineração de Dados na Agricultura**

A MD pode ser aplicada em várias áreas do conhecimento, entre elas, a medicina, marketing, educação, bioinformática e também na área agronômica, que é o foco deste trabalho.

Por se tratar de um campo de pesquisa novo, ainda não há nenhuma técnica de MD voltada especificamente para solucionar problemas da área agronômica (MUCHERINO *et al.*, 2009). Apesar disso, os autores afirmam que a MD pode ser uma técnica eficiente para resolução de problemas agrícolas complexos.

A MD com o objetivo de classificação tem sido usada em diferentes contextos agronômicos como pode ser visto em Souza *et al.* (2010) que avaliaram os atributos químicos do solo e a produtividade da cultura de cana-de-açúcar utilizando técnicas da geoestatística e MD pela indução de Árvore de Decisão. O estudo concluiu que a altitude e K são os atributos com as maiores correlações com a produtividade de cana-de-açúcar, além disso, a utilização da técnica de Árvore de Decisão permitiu verificar que o atributo altitude apresenta maior potencial para interpretar mapas de produtividade de cana-de-açúcar.

Outro trabalho utilizando a técnica de Árvore de Decisão foi feito visando o auxílio na compreensão de manifestações epidêmicas da ferrugem do cafeeiro (MEIRA *et al.*, 2008). Este trabalhou também apresentou bons resultados, de modo que foi possível auxiliar na compreensão de quais atributos e como as interações destes atributos conduziram as epidemias da ferrugem no campo cafeeiro.

Com foco na agrometeorologia, Meira e Rodrigues (2003) utilizaram a MD para analisar uma base com dados agro-meteorológicos visando descobrir regras de associação temporal relacionadas ao surgimento de doenças na cultura do café. Também, Mathias (2006) desenvolveu um sistema baseado em Redes Neurais Artificiais (RNA) para o tratamento de dados climáticos visando o reconhecimento de padrões dos atributos agro-meteorológicos em relação ao molhamento foliar pelo orvalho.

Como afirmam Silva *et al.* (2009), a variabilidade climática deve ser analisada, pois os sistemas agrícolas podem ser afetados com sérias consequências para a produção de alimentos. Bucene *et al.* (2002), concordam, pois afirmam que ocorrem grande perdas de produtividade devido a ocorrência de sinistros relacionados a eventos climáticos. O trabalho de Boschi *et al.* (2011) utilizou a MD para analisar o comportamento espaço-temporal da precipitação pluvial no Estado do Rio Grande do Sul. Segundo Boschi *et al.* (2011) o uso das técnicas de MD é uma alternativa promissora por ser capaz de transformar dados em informações e, posteriormente, em conhecimento. Como resultado tem-se que a tarefa de agrupamento de MD mostrou-se eficiente na análise do comportamento espaço-temporal da precipitação pluviométrica.

Seguindo a importância de estudos que avaliam os atributos de precipitação pluvial, tem-se o trabalho de Dourado *et al.* (2012) cujo objetivo foi identificar e analisar as zonas pluviométricas homogêneas no Estado da Bahia, para isso foi utilizado o algoritmo *K-means* para segmentação dos dados, de modo que foram identificados quatro zonas homogêneas no Estado.

Concordando com o trabalho de Boschi *et al.* (2011), Dourado *et al.* (2012) também afirmam que a tarefa de agrupamento foi capaz de identificar regiões homogeneamente pluviométricas, mostrando-se eficiente para o objetivo proposto.

Bucene *et al.* (2002) propôs o uso de MD para identificar relações entre atributos climáticos como temperatura máxima, temperatura mínima, precipitação entre outros, para descobrir eventuais novos acontecimentos que possibilitem realizar a previsão de geada para a cultura do café e previsão do *défict* hídrico para as culturas da cana-de-açúcar e café no Estado de São Paulo.

Ainda envolvendo a tarefa de classificação de MD, Guimarães *et al.* (2002) realizaram um estudo para determinar o comportamento da produtividade de uma safra em função das propriedades físico-químicas do solo, utilizando a técnica de Algoritmos Genéticos. A partir desse estudo, Guimarães (2005), utilizou a técnica de Algoritmos

Genéticos para desenvolver uma nova metodologia baseada em Algoritmos Genéticos capaz de analisar os dados físico-químicos da água e do solo, a qual se mostrou eficiente como ferramenta auxiliar na compreensão da variabilidade dos dados. Outro estudo envolvendo os atributos físico-químicos do solo foi o de Vriesmann *et al.* (2004), onde foram aplicadas as técnicas de Árvore de Decisão e Algoritmos Genéticos na análise de características físico-químicas do solo associadas a produtividade da soja.

Outra aplicação da MD no campo da agricultura está na previsão de mapas de solo (BEHRENS e SCHOLTEN, 2006). Os autores afirmam que em se tratando de ciência do solo, as técnicas de MD são capazes de aprender a relação entre as classes de solo mapeadas bem como fatores de sua formação, sendo possível assim construir automaticamente mapas digitais do solo de modo que auxilie o mapeamento de campo. As técnicas de MD utilizadas no trabalho foram: RNAs, Máquina de Vetor de Suporte (SVM, do inglês *Support Vector Machine*), e classificação baseada em Árvores de Decisão. Os autores concluíram que a RNA foi a melhor técnica para prever classes de solo, ao contrário da técnica SVM e Árvore de Decisão, cujo tempo computacional foi lento.

Ainda envolvendo a tarefa de classificação, Antunes *et al.* (2011) desenvolveram um estudo visando classificar as fases fenológicas da cultura da cana-de-açúcar no Estado de São Paulo. O resultado desta pesquisa foi satisfatório, tendo os autores confirmando a aderência das técnicas de MD para o objetivo proposto.

Quanto a tarefa de regressão, pode ser visto o trabalho de Molin *et al.* (2001), em que foram aplicados métodos de regressão linear múltipla sobre o rendimento de grãos e as possíveis causas da variabilidade de produção relacionada à fertilidade do solo utilizando dados de uma área sendo cultivada sob SPD há anos.

Quanto à tarefa de classificação, há o trabalho de Weber *et al.* (2011), que com a utilização da técnica RNA, objetivou avaliar a influência da compactação do solo na produtividade através das medidas de correlação de Pearson e Erro Médio Absoluto (EMA) obtidas a partir da aplicação da RNA. Constatando-se que o atributo de índice de compactação do solo pode auxiliar na construção de modelos que objetivem estimar a produtividade da cultura do milho, e, além disso, que este atributo não pode ser estudado isoladamente, sendo necessária sua análise em conjunto com os atributos químicos do solo.

Outro trabalho que pode ser citado é o de Uno *et al.* (2005), que aliado a estatística também utilizou RNA para desenvolver modelos de previsão de rendimento a

partir de vários índices de vegetação. No entanto, os autores afirmam que, apesar do potencial da RNA, em particular na criação de mapas de produtividade, é necessário mais estudos antes da aplicação em campo.

## **2.6. Mineração de Dados em Pequenas Bases**

No domínio da agricultura, os experimentos comumente envolvem um grande número de atributos e um número de observações variando em torno de 50 a 500 pontos amostrados. Isso ocorre devido à grande dificuldade que existe na instalação e acompanhamento dos experimentos, o que inviabiliza uma grande quantidade de amostras. É o caso do trabalho de Souza *et al.* (2010), o qual envolveu 118 pontos amostrais.

Outro trabalho que pode ser citado como exemplo é o de Silva *et al.* (2004), cujo objetivo foi utilizar RNA, por meio de 441 registros, para a determinação de necessidades de adubação da cultura da goiabeira.

Conceitualmente, a MD destina-se a análise de uma grande base de dados, entendendo-se como “grande”, milhares a milhões de registros. Porém, a complexidade existente na análise de pequenas bases em função do número de atributos e de suas características, associada à busca por novos conhecimentos nas mesmas, tem levado a estudos voltados para a MD em pequenas bases, buscando correlações com característica de novidade.

Esse pensamento pode ser visto no trabalho de Williams *et al.* (2002), que sugeriu utilizar os métodos estatísticos clássicos quando a base de dados for considerada pequena. No entanto, a idéia de utilizar MD consiste na busca por algum conhecimento, além de confiável, também inesperado.

A MD aplicada em pequenos conjuntos de dados é um campo de pesquisa relativamente novo, e já existindo técnicas capazes de tratar com essa característica. Do ponto de vista computacional, a análise de pequenas bases de dados é um ponto importante no Aprendizado de Máquina uma vez que com poucos dados a tendência é ter informações incompletas (LI e LIU, 2010). Li e Yeh (2008) afirmam ainda que, o tamanho das amostras pode influenciar de modo significativo na precisão de aprendizagem.

Existe uma discussão sobre a quantificação de uma pequena amostra. Alguns trabalhos encontrados na literatura indicam a quantidade de amostras que caracterizam uma pequena base de dados. Huang (2002), que estuda o princípio de difusão da

informação com aplicação da teoria dos conjuntos fuzzy, em sua pesquisa, afirma que uma amostra pequena é o equivalente a menos de 30 registros. O autor afirma ainda que, uma base de dados é pequena se o seu tamanho é muito pequeno para fornecer a precisão necessária para obter o reconhecimento de uma relação da qual a base foi projetada.

Outro trabalho que indica a quantidade de amostras que foram analisadas como uma pequena base de dados é o de Yang e Kacman (2009), que propõem um algoritmo hiperplano local adaptativo para trabalhar com pequenas bases, e os dados trabalhados nesse estudo envolvem aplicação na área da medicina. Nesse trabalho a quantidade de amostras é de 699 registros referentes a um conjunto de dados sobre o câncer de mama, sendo 16 registros faltantes, 444 registros para a classe ‘benigno’ e 239 registros para a classe ‘maligno’.

Pode-se citar também o trabalho de Ng e Han (1994), cujo objetivo foi explorar o papel da tarefa de clusterização na Mineração Espacial. Ng e Han (1994) executaram a tarefa de agrupamento utilizando o algoritmo *Partitioning Around Medoids* (PAM) e concluíram que o mesmo funcionou satisfatoriamente em pequenos conjuntos, em que se assume como pequeno, por exemplo, 100 amostras e 5 *clusters*.

Com isso, percebe-se que não há um número exato que possa ser caracterizado como pequeno. Li e Liu (2010), afirmam que um conjunto de dados “pequeno” é um conceito relativo. Pode ser visto que alguns autores adotam um número de amostras e o caracterizam como pequeno ou grande conforme seu entendimento.

Mesmo com a discussão envolvendo o tamanho de uma base de dados caracterizada como “pequena”, é possível encontrar na literatura trabalhos que propõem uma abordagem para se trabalhar com bases contendo tal característica. É o caso do trabalho de Li e Liu (2010), que propuseram uma nova abordagem de construção de atributos para estender pequenas bases de dados visando aumentar a precisão de classificação. De forma diferente à geração de amostras virtuais, este novo método proposto além de evitar uma tendência na estrutura de dados devido à geração de amostras virtuais, também aumenta a informação do dado. O principal objetivo desse trabalho não foi simplesmente aumentar a quantidade de dados, mas aumentar a dimensão do dado (informação), aplicando-se para isso a função de pertinência fuzzy.

No trabalho sobre o ambiente de manufatura de Li e Yeh (2008), os dados são escassos, pois são obtidos apenas nos estágios iniciais do sistema de manufatura. Desse

modo, a pesquisa propõe um algoritmo de aprendizado não paramétrico<sup>3</sup> para pequenos conjuntos de dados ao invés de utilizar técnicas de estatística paramétrica. Huang e Moraga (2004) apresentam outra técnica para trabalhar com pequenas amostras, em que utilizam a combinação de RNAs convencionais e a técnica de difusão de informação, que usa teoria fuzzy para obter novas amostras. Dessa forma, sugere-se uma Rede Neural Difusa (DNN, do inglês *Diffusion Neural Network*), que objetiva minimizar o erro e trabalhar em bases pequenas.

Huang *et al.* (2006) dizem que RNAs convencionais tem baixo desempenho quando há atributos irrelevantes ou em bases pequenas, ao contrário outras técnicas como a programação genética. No trabalho de Huang *et al.* (2006) é proposto o *two stage genetic programming* (2SGP), que combina as vantagens das regras SE-ENTÃO e a função discriminante. As regras são derivadas usando programação genética, em seguida, o número reduzido de dados é alimentado com a programação genética para formar a função discriminante para fornecer a capacidade de previsão. Os autores afirmam que, a razão de utilizar programação genética é que esta técnica pode, automaticamente e heurísticamente, determinar a função discriminante, além de trabalhar com pequenas bases de dados. Os resultados dessa pesquisa comprovam que o método proposto mostrou-se eficiente tanto para grandes quanto para pequenas bases de dados.

O tamanho das amostras é um fator que pode influenciar no desempenho do aprendizado, e ainda segundo Li e Liu (2009), uma proposta que vem sendo estudada é a geração de dados virtuais para resolver o problema de dados incompletos. Como no trabalho proposto por Zhu *et al.* (2006), em que utilizando técnicas para criação de novas amostras, buscando preencher lacunas existentes na base e com a utilização de máquina de vetor de suporte, para obtenção da probabilidade posterior, um novo método de aprendizagem de RNA baseada em probabilidade posterior foi desenvolvido, e com resposta positiva, uma vez que a precisão da aprendizagem foi melhorada por meio do algoritmo proposto.

Outro trabalho envolvendo o uso de técnicas para gerar amostras virtuais é proposto por Tsai e Li (2008), cujo objetivo é utilizar a técnica *bootstrap* para preencher as lacunas de informação de dados esparsos. Essa pesquisa também utilizou a técnica de

---

<sup>3</sup> Ausência de conhecimento *a priori*

RNA como forma de extração de conhecimento, e obteve um resultado positivo, já que a taxa de erro de predição foi significativamente reduzida através do método proposto.

Tsai e Li (2008) afirmam que estudos envolvendo bases pequenas são poucos até o momento e existe um grande potencial para pesquisas nessa área visando melhorar a taxa de precisão no processo de aprendizagem.

## **2.7. Pré-processamento**

### **2.7.1. Seleção de atributos**

A técnica de MD não consiste apenas em definir a tarefa a ser utilizada e aplicar os algoritmos existentes. É de grande importância o tratamento correto dos dados dentro da fase de pré-processamento, para isso, existem diversas técnicas que objetivam melhorar o resultado da MD.

A seleção de atributos é uma das principais tarefas de pré-processamento utilizadas para preparar os dados que serão minerados posteriormente. Essa seleção tem como objetivo, (i) melhorar o desempenho do algoritmo de MD, (ii) remover dados com ruídos, que são aqueles que contêm erros ou estão fora da faixa aceitável de valores (*outliers*), e (iii) reduzir a dimensionalidade dos dados (quantidade de atributos), de modo que sejam selecionados apenas os atributos mais relevantes sem que ocorra a perda de informação.

Nos casos de bases de dados de alta dimensionalidade pode ocorrer a existência de atributos redundantes, irrelevantes ou ainda, com pouca influência sobre o atributo de interesse (atributo meta). Outro problema referente à alta dimensionalidade dos dados é que esta impõe exigência sobre o tamanho da amostra que deve ser usada na indução de modelos. Jolliffe (1972) afirma que bases de dados contendo dez ou mais atributos já podem ser considerados base de dados com alta dimensionalidade.

Desse modo, uma possível abordagem para diminuir esta exigência, é através de técnicas de seleção de atributos para reduzir a dimensionalidade da base de dados. As técnicas de seleção de atributos são empregadas antes da geração de modelos com o objetivo de selecionar apenas os atributos mais relevantes para a descrição do objeto de estudo (PIRAMUTHU, 2004).

A seleção de atributos pode ser muito útil quando se tem uma base de dados com muitos atributos e poucos registros (PAPPA, 2002). Nesse caso, é importante que seja aplicada alguma técnica de seleção de atributos, pois há benefícios na redução da dimensionalidade da base de dados, entre eles, está o fato de que muitos algoritmos de

MD funcionam melhor se a dimensionalidade dos dados for menor. Isso ocorre porque a redução de dimensionalidade é capaz de eliminar atributos irrelevantes que interfeririam negativamente no resultado, podendo levar a um modelo mais compreensível pelo fato de envolver menos atributos (TAN *et al.*, 2009).

A evolução da computação, como por exemplo, no desenvolvimento de algoritmos mais eficientes para manipulação de imagens digitais, tem possibilitado um número crescente de aplicações de reconhecimento de padrões em sensoriamento remoto (OLIVEIRA *et al.*, 2005). Nesse sentido, os trabalhos de Santos *et al.* (2007) e Oliveira *et al.* (2005) são exemplos de estudos que utilizam a seleção de atributos para classificação de regiões. Em ambos os trabalhos recomenda-se a utilização dessa técnica, mostrando-se eficiente no seu propósito.

Algumas das abordagens mais comuns para redução da dimensionalidade, especialmente para dados contínuos, utilizam técnicas de álgebra linear para projetar os dados de um espaço de alta dimensionalidade para um de dimensionalidade menor (TAN *et al.*, 2009), como por exemplo, a seleção de atributos baseada na técnica de Análise de Componentes Principais (ACP).

#### 2.7.1.1. Análise de componentes principais (ACP)

A ACP é uma técnica para seleção de atributos que transforma o conjunto de dados originais, tal que os dados transformados são projetados em um espaço que maximiza a variância da amostra (JOLLIFFE, 2002). A ACP determina um novo conjunto de atributos, que são chamados de componentes principais e é uma combinação linear dos atributos originais. As combinações lineares são ortogonais entre si e mantendo o máximo possível da variação de dados. Os componentes principais podem ser calculados empregando-se a matriz de covariância amostral  $S$  usando a fórmula:

Equação 1 – Matriz de covariância (Ferreira, 2008).

$$S = \begin{pmatrix} cov(X_1, X_1) & \dots & cov(X_1, X_p) \\ \dots & \dots & \dots \\ cov(X_p, X_1) & \dots & cov(X_p, X_p) \end{pmatrix}, \text{ onde } cov(X_i, X_j) \text{ é a covariância de } X_i \text{ e } X_j$$

Seja  $\mathbf{X}_{np}$  uma base de dados com  $p$  variáveis e  $n$  amostras, e seja  $\{X_1 \dots X_p\}$  o conjunto de variáveis originais. A ACP determina um novo conjunto de variáveis  $\{Y_1 \dots$

$Y_p\}$  em que cada  $Y_i$  é chamado de componente principal e é uma combinação linear das variáveis originais.

O principal objetivo da ACP é reduzir a dimensionalidade dos dados multivariados de modo que se preserve o máximo possível de informação contida nos mesmos (RODRIGUES, 2011).

Segundo Tan *et al.* (2009), dentre algumas características interessantes da ACP pode-se citar, o fato de que a ACP tende a identificar padrões mais fortes nos dados, desse modo, pode ser utilizada também como técnica de descoberta de padrões. Outra característica é que a variabilidade dos dados pode ser capturada por uma pequena fração do conjunto total de atributos, assim é possível aplicar técnicas de MD que não funcionam muito bem em dados de alta dimensionalidade. Também é possível a eliminação dos ruídos, pois, na maioria dos casos, os ruídos presentes nos dados são mais fracos que os padrões, sendo assim, a redução de dimensionalidade é capaz de eliminar muito desse ruído.

A ACP é uma das técnicas mais utilizadas para seleção de atributos, e na área agrônômica encontram-se trabalhos envolvendo esta técnica em diferentes domínios de aplicação. Valladares *et al.* (2008) utilizaram ACP para agrupar os perfis de Organossolos quanto às suas semelhanças com base em atributos morfológicos, físicos e químicos, e assim estudar suas correlações. Os resultados permitiram concluir que a técnica mostrou-se adequada para o objetivo proposto.

Guarienti *et al.* (2003) aplicaram a ACP sobre uma base de dados contendo atributos meteorológicos, para verificar quais eram os que mais afetavam a qualidade industrial e o rendimento de grãos do trigo nas condições da região tritícola do sul do Brasil.

Sodré *et al.* (2007) tinham o objetivo verificar as características físico-químicas de amostras de mel de *Apis mellifera* L. (Hymenoptera: Apidae) coletadas em cidades do Estado do Ceará e, para isso, utilizaram a ACP para avaliar a importância de cada característica físico-química estudada sobre a variação total disponível.

Com o objetivo de analisar a variabilidade espacial de atributos químicos de um Latossolo Vermelho-Amarelo húmico cultivado com café, Silva *et al.* (2010) empregaram a ACP como uma técnica complementar. Os dados foram analisados por meio da estatística descritiva, seguida da ACP e por fim uma análise de agrupamento.

Para a realização da ACP há alguns critérios para a seleção dos atributos, entre eles estão os critérios B2 e B4 (JOLLIFFE, 1972). O critério B2 é realizado executando-

se a ACP sobre um conjunto de dados com todos os atributos originais, então se inspecionam os  $k$  autovalores de modo que não ultrapassem um limiar  $\lambda_0$ , retêm-se os  $k$  autovetores correspondentes aos  $k$  autovalores. Assim, analisa-se o autovetor de menor autovalor correspondente, então, analisa-se o autovetor com segundo menor autovalor, e assim sucessivamente até analisar os  $k$  autovetores. Em seguida, escolhe-se um atributo com o maior coeficiente naquele componente (autovetor) que está em análise considerando que o atributo selecionado ainda não foi associado aos componentes analisados anteriormente. Então os  $k$  atributos selecionados são removidos da base de dados. O critério B4 é semelhante ao critério B2, sendo que a diferença é que o B4 analisa os  $k$  componentes do primeiro para o último. Assim, os métodos B2 e B4 retêm  $p-k$  atributos originais, eliminando os  $k$  atributos que menos contribuem para explicação da variação dos dados (OLIVEIRA Jr., 2012).

Outro critério que pode ser utilizado é a Análise de Componentes Principais Supervisionada (ACPS) proposta por Bair *et al.* (2006). Como a ACP é um processo não supervisionado, em que não se considera a influência dos atributos de entrada sobre o atributo meta, a ACPS concentra-se apenas nos atributos que possuem maior influência no comportamento do atributo meta. Oliveira Jr. (2012) explica que, nesse caso, a informação fornecida pelo atributo meta é utilizada como um pré-filtro sobre os atributos de entrada da ACP, assim, o pré-filtro seleciona apenas os atributos que têm associação mais forte com o atributo meta e a ACP é aplicada no conjunto de atributos pré-selecionados. Oliveira Jr. (2012) realizou um estudo em que foi demonstrado que os melhores regressores e classificadores foram gerados com o conjunto de atributos selecionados pela ACPS.

Contudo a abordagem apresentada por Joliffe (1972) e Bair *et al.* (2006) não exploram o conhecimento sobre o domínio da aplicação quando da seleção de atributos, e a incorporação de conhecimento *a priori* pode ter um grande impacto sobre os resultados da MD (SARABJOT *et al.* 1995; RUSSELL e NORVIG, 2004).

Alguns trabalhos abordam o uso de informação do domínio no cômputo da ACP, por exemplo, pela suposição de restrições de independência, expressas em um modelo gráfico gaussiano, sobre a inversa da matriz de covariância (WIESEL e HERO, 2009), também o trabalho de Bürkle (2006), que propõe uma nova abordagem de poda semântica de regras de associação que se baseia no uso de conhecimento *a priori* sobre o domínio, representado pelas relações de dependência entre os atributos. Contudo,

poucos trabalhos têm explorado o uso de independência marginal<sup>4</sup> na seleção de atributos usando ACP, técnica esta que é importante para a simplificação de modelos (SILVA e GHAHRAMANI, 2009).

Cramer (1999) propõe uma forma de estimar a matriz de covariância quando existe algum conhecimento, sobre uma hipótese de independência marginal entre determinados atributos da base de dados. O conhecimento *a priori* sobre o domínio baseado em hipótese de independência marginal pode ser uma forma interessante de se realizar a seleção de atributos, e assim, selecionar os melhores subconjuntos de atributos para a análise das bases de dados obtendo-se modelos de MD com melhores resultados.

## **2.7.2. Transformação das bases de dados**

### **2.7.2.1. SMOTE**

*Synthetic Minority Over-Sampling Technique* (SMOTE) é uma técnica de desbalanceamento de dados que ao invés de diminuir os elementos da classe majoritária, a técnica aumenta os elementos da classe minoritária. Desse modo, a técnica acaba gerando novos dados sintéticos (ALMEIDA e RODRIGUES, 2011). Esses novos dados sintéticos são gerados na vizinhança de cada caso da classe minoritária, de forma a fazer crescer a região de decisão, de modo a aumentar o poder de generalização dos classificadores gerados para esses dados (MACHADO, 2007).

Machado (2007) explica que, no espaço amostral do conjunto de dados, esses novos dados sintéticos serão interpolados aleatoriamente ao longo do segmento de reta que liga cada caso da classe minoritária a um de seus  $k$  vizinhos mais próximos, escolhidos de forma aleatória.

### **2.7.2.2. Método mega-trend-diffusion**

Uma alternativa que pode ser aliada à seleção de atributos refere-se as técnicas de ampliação de base de dados, em que é possível a geração de amostras virtuais baseadas nos dados reais visando o aumento da quantidade de registros.

Os trabalhos que foram citados referentes à MD em pequenas bases, em que alguns destes objetivavam a criação ou alteração de algoritmos capazes de criar essas amostras virtuais também podem ser classificados como técnicas de pré-processamento.

---

<sup>4</sup> Independência marginal: X e Y são independentes quando se ignora a informação sobre Z

O trabalho de Li *et al.* (2007a), reconhece que ter dados suficientes nem sempre é possível, por esse motivo os autores utilizaram técnicas de expansão de dados em pequenas bases visando melhorar a precisão de aprendizado na área de manufatura. Nessa pesquisa os autores propuseram o método *mega-trend-diffusion* (MTD) para estimar a faixa de domínio de um pequeno conjunto de dados e produzir amostras artificiais para o treinamento de uma RNA com o algoritmo *backpropagation* modificado (BPNN). O resultado desse trabalho foi significativo, pois mostrou bom desempenho em pequenas bases de dados.

Li *et al.* (2007b) aplicaram o método MTD em uma pequena base de dados, cujo domínio é o diagnóstico de câncer com o objetivo de melhorar a precisão do diagnóstico do gene de câncer de bexiga. Os resultados obtidos foram satisfatórios, os modelos demonstraram que quando a quantidade de dados foi aumentada com o método MTD a precisão de aprendizagem do diagnóstico de câncer de bexiga aumentou de forma estável de 81,67% para 100%.

Em 2011, Liao propôs o uso das técnicas de seleção de atributos para trabalhar com pequenas de dados envolvendo dados sobre câncer de bexiga, visando demonstrar que essas técnicas são melhores do que as abordagens existentes, como o método MTD proposto por Li *et al.* (2007a), em se tratando do domínio de predição do diagnóstico de câncer de bexiga. Esse trabalho propôs a redução da dimensionalidade dos dados através da seleção de atributos, de modo que se obteve resultado igual ou superior as abordagens mencionadas.

Porém, outro trabalho de Li *et al.* (2010) reforça os resultados significativos do método MTD, em que se aplicou o método MTD sobre conjuntos de dados médicos desbalanceados. Esse problema pode ser exemplificado da seguinte forma, sejam os valores do atributo meta C1, em que esse valor representa os pacientes que tem uma determinada doença e C2 representando os pacientes que não tem a doença, tem-se  $C1 = 690$  e  $C2 = 10$ .

Isso é considerado um problema, pois pode ocorrer de o algoritmo classificar corretamente 650 pacientes que tem a doença e não classificar corretamente nenhum registro cujo paciente não tem a doença, ou seja, apesar do classificador ter uma alta acurácia para C1, o resultado é ruim para C2. Desse modo, existem técnicas capazes de tratar o desbalanceamento das classes.

Para trabalhar com o desequilíbrio de classes, Li *et al.* (2010) optou por reduzir a quantidade de registros quando se tem muitas amostras para uma determinada classe

(classe majoritária) e aumentar a quantidade de registros para a classe com poucas amostras (classe minoritária).

Com isso, foi utilizado o método MTD para realizar a criação de amostras virtuais para a classe minoritária. Os autores concluíram que esta foi uma abordagem útil para tratar o problema de desbalanceamento de classes, principalmente em se tratando de dados da área médica em que os autores afirmam ser comum esse tipo de problema nas bases de dados.

## **2.8. Pós-processamento**

### **2.8.1. Interpretação dos resultados**

Essa etapa assegura que apenas os resultados válidos e úteis sejam incorporados ao sistema de apoio de decisões (TAN *et al.*, 2009). O pós-processamento é a última etapa do processo KDD, envolvendo a visualização, a análise e interpretação dos resultados obtidos na etapa de MD.

Dentre as atividades que podem compor esta etapa estão as medições estatísticas ou métodos de hipóteses, de modo que eliminem os resultados não legítimos obtidos na etapa de MD (TAN *et al.*, 2009). Outra atividade que pode ser considerada parte da etapa de pós-processamento é a consolidação resultados, em que é possível a verificação do novo conhecimento sobre novos dados.

As novas informações descobertas, que também são chamadas de padrões, devem possuir algum grau de certeza para que sejam consideradas válidas, e precisam ser descritas em uma linguagem compreensível para o usuário para que seja possível realizar uma análise mais profunda (PIMENTEL e OMAR, 2006). Por essa razão é que o uso da MD requer uma interação muito forte com especialista humano da área de domínio, pois são eles que podem determinar o grau de importância do conhecimento descoberto (NAVEGA, 2002; PIMENTEL e OMAR, 2006; CASTANHEIRA, 2008).

### 3. METODOLOGIA

#### 3.1. Áreas de estudo

O estudo foi realizado em três áreas experimentais localizadas na região dos Campos Gerais do Paraná.

A primeira área, denominada área A, está situada no município de Ponta Grossa (25°8' S, 50°15' W, e altitude média de 853 m), em um Latossolo Vermelho textura média (295 g kg<sup>-1</sup> de argila) com alta acidez. A área se encontrava há 15 anos sob SPD quando o experimento foi instalado. Aplicaram-se quatro doses de gesso agrícola (0, 4, 8 e 12 t ha<sup>-1</sup>) à lanço sobre a superfície do solo em 1993, em um delineamento de blocos completos ao acaso com três repetições. Amostras de solo foram coletadas nas camadas de 0–10 (A), 10–20 (B), 20–40 (C) e 40–60 (D) cm, aos 8, 20, 32 e 44 meses após a aplicação de gesso, durante o período de 1994 a 1998.

A segunda área, denominada área B, também está localizada no município de Ponta Grossa (25°10' S, 50°05' W, e altitude média de 970 m), em um Latossolo Vermelho textura argilosa (580 g kg<sup>-1</sup> de argila) com acidez média, até então utilizado para pastagem. Aplicaram-se quatro doses de gesso agrícola (0, 3, 6 e 9 t ha<sup>-1</sup>) à lanço sobre a superfície do solo em 1998, em um delineamento de blocos completos ao acaso com três repetições. Amostras de solo foram coletadas nas camadas de 0–10 (A), 10–20 (B), 20–40 (C) e 40–60 (D) cm, aos 8, 20, 32, 44, 56, 68, 80 e 92 meses após a aplicação de gesso, durante o período de 1999 a 2006.

A terceira área, denominada área C, está localizada no município de Guarapuava (25°17' S, 51°48' W, e altitude média de 1080 m), em um Latossolo Vermelho textura argilosa (650 g kg<sup>-1</sup> de argila) com baixa acidez. A área se encontrava há 15 anos sob SPD quando o experimento foi instalado. Aplicaram-se quatro doses de gesso agrícola (0, 4, 8 e 12 t ha<sup>-1</sup>) à lanço sobre a superfície do solo em 2005, em um delineamento de blocos completos ao acaso com quatro repetições. Amostras de solo foram coletadas nas camadas de 0–10 (A), 10–20 (B), 20–40 (C) e 40–60 (D) cm, aos 8 e 20 meses após a aplicação de gesso, durante o período de 2006 a 2008.

De acordo com o sistema Köppen-Geiger (PEEL *et al.*, 2007), o clima das três áreas experimentais é classificado como Cfb, com verões frescos e geadas frequentes durante o inverno. A temperatura média anual é de 17°C, sendo a temperatura média do mês mais frio de 13°C e a do mês mais quente não superior a 22°C. A precipitação pluvial média anual é de 1600 mm, com chuvas concentradas entre os meses de setembro a abril.

As três áreas experimentais foram manejadas no SPD com rotação de culturas, sendo cultivados milho (*Zea mays* L.) ou soja (*Glycine max* L. Merrill) durante a estação de primavera-verão, e aveia preta (*Avena strigosa* Schreb.) ou trigo (*Triticum aestivum* L.) ou cevada (*Hordeum vulgare* L.) na estação de outono-inverno. Fertilizantes foram aplicados de acordo com as necessidades das culturas e os teores de nutrientes no solo, conforme as recomendações para o Estado do Paraná. Maiores detalhes das áreas experimentais utilizadas nesse estudo podem ser vistos nas publicações de Caires *et al.* (1999, 2001, 2002, 2004, 2011).

As análises químicas de solo, realizadas antes da instalação do experimento, para as três áreas de estudo são apresentadas na Tabela 2.

Tabela 2 – Análises químicas do solo para as três áreas de estudo

| <b>Área A</b>          |                                  |                                                  |                             |                                                  |                                                  |                                                 |                                           |
|------------------------|----------------------------------|--------------------------------------------------|-----------------------------|--------------------------------------------------|--------------------------------------------------|-------------------------------------------------|-------------------------------------------|
| <b>Camada<br/>(cm)</b> | <b>pH<br/>(CaCl<sub>2</sub>)</b> | <b>Al<br/>(mmol<sub>c</sub> dm<sup>-3</sup>)</b> | <b>Saturação<br/>por Al</b> | <b>Ca<br/>(mmol<sub>c</sub> dm<sup>-3</sup>)</b> | <b>Mg<br/>(mmol<sub>c</sub> dm<sup>-3</sup>)</b> | <b>K<br/>(mmol<sub>c</sub> dm<sup>-3</sup>)</b> | <b>C-orgânico<br/>(g dm<sup>-3</sup>)</b> |
| <b>0 a 20</b>          | 4,5                              | 6                                                | 18%                         | 16                                               | 10                                               | 1,4                                             | 19                                        |
| <b>20 a 40</b>         | 4,4                              | 6                                                | 18%                         | 16                                               | 10                                               | 1,5                                             | 18                                        |
| <b>Área B</b>          |                                  |                                                  |                             |                                                  |                                                  |                                                 |                                           |
| <b>Camada<br/>(cm)</b> | <b>pH<br/>(CaCl<sub>2</sub>)</b> | <b>Al<br/>(mmol<sub>c</sub> dm<sup>-3</sup>)</b> | <b>Saturação<br/>por Al</b> | <b>Ca<br/>(mmol<sub>c</sub> dm<sup>-3</sup>)</b> | <b>Mg<br/>(mmol<sub>c</sub> dm<sup>-3</sup>)</b> | <b>K<br/>(mmol<sub>c</sub> dm<sup>-3</sup>)</b> | <b>C-orgânico<br/>(g dm<sup>-3</sup>)</b> |
| <b>0 a 20</b>          | 4,6                              | 3                                                | 6%                          | 25                                               | 20                                               | 3,6                                             | 31                                        |
| <b>20 a 40</b>         | 4,2                              | 8                                                | 32%                         | 7                                                | 8                                                | 2,2                                             | 21                                        |
| <b>Área C</b>          |                                  |                                                  |                             |                                                  |                                                  |                                                 |                                           |
| <b>Camada<br/>(cm)</b> | <b>pH<br/>(CaCl<sub>2</sub>)</b> | <b>Al<br/>(mmol<sub>c</sub> dm<sup>-3</sup>)</b> | <b>Saturação<br/>por Al</b> | <b>Ca<br/>(mmol<sub>c</sub> dm<sup>-3</sup>)</b> | <b>Mg<br/>(mmol<sub>c</sub> dm<sup>-3</sup>)</b> | <b>K<br/>(mmol<sub>c</sub> dm<sup>-3</sup>)</b> | <b>C-orgânico<br/>(g dm<sup>-3</sup>)</b> |
| <b>0 a 20</b>          | 6,0                              | Nd                                               | -                           | 48                                               | 22                                               | 4,2                                             | 28                                        |
| <b>20 a 40</b>         | 5,2                              | Nd                                               | -                           | 23                                               | 29                                               | 2,4                                             | 20                                        |

### 3.2. Sobre os atributos

A partir das amostras do solo coletadas nas diferentes profundidades e épocas de amostragem, determinaram-se os teores de Al, Ca, Mg e K trocáveis, de acordo com métodos padrão usados pelo Instituto Agronômico do Paraná (PAVAN *et al.*, 1992). A capacidade de troca catiônica efetiva (CTCe) foi calculada pela soma de cátions trocáveis (Al + Ca + Mg + K), e a saturação por Al, Ca, Mg e K na CTCe foi calculada pelas equações:

$$\text{Saturação por Al (\%)} = (\text{Al}/\text{CTCe}) \times 100$$

$$\text{Saturação por Ca (\%)} = (\text{Ca}/\text{CTCe}) \times 100$$

$$\text{Saturação por Mg (\%)} = (\text{Mg}/\text{CTCe}) \times 100$$

$$\text{Saturação por K (\%)} = (\text{K}/\text{CTCe}) \times 100$$

Desse modo, a partir das análises de solo, trabalhou-se com os teores de Al, Ca, Mg e K trocáveis, em  $\text{mmol}_c \text{ dm}^{-3}$ , e com os valores de saturação por Al (m), saturação por Ca, saturação por Mg e saturação por K na CTCe, em percentagem (%). É importante destacar que, além dos atributos químicos do solo, trabalhou-se também com o atributo época, que representa o tempo, em meses, após a aplicação do gesso agrícola, e por fim, com o atributo gesso, que representa a dose de gesso aplicada em  $\text{t ha}^{-1}$ .

Todos os atributos químicos do solo foram separados por camada (A, B, C e D), obtendo-se, no total, 34 atributos:  $\text{Al}_A$ ,  $\text{Al}_B$ ,  $\text{Al}_C$ ,  $\text{Al}_D$ ,  $\text{Ca}_A$ ,  $\text{Ca}_B$ ,  $\text{Ca}_C$ ,  $\text{Ca}_D$ ,  $\text{Mg}_A$ ,  $\text{Mg}_B$ ,  $\text{Mg}_C$ ,  $\text{Mg}_D$ ,  $\text{K}_A$ ,  $\text{K}_B$ ,  $\text{K}_C$ ,  $\text{K}_D$ ,  $m_A$ ,  $m_B$ ,  $m_C$ ,  $m_D$ ,  $\text{CaCTCe}_A$ ,  $\text{CaCTCe}_B$ ,  $\text{CaCTCe}_C$ ,  $\text{CaCTCe}_D$ ,  $\text{MgCTCe}_A$ ,  $\text{MgCTCe}_B$ ,  $\text{MgCTCe}_C$ ,  $\text{MgCTCe}_D$ ,  $\text{KCTCe}_A$ ,  $\text{KCTCe}_B$ ,  $\text{KCTCe}_C$ ,  $\text{KCTCe}_D$ , época e gesso.

### 3.3. Organização das bases de dados

O período da amostragem de solo foi diferente em cada área experimental, tendo sido de até 44, 92 e 20 meses após a aplicação de gesso, respectivamente nas áreas A, B e C. Para que a análise dos dados fosse realizada de modo consistente e confiável, as bases de dados foram unidas de acordo com o tempo, em meses após a aplicação de gesso (atributo época), de modo que o atributo tivesse a mesma quantidade de meses após a aplicação de gesso. Assim, houve a possibilidade de realizar a junção dos dados das áreas A, B e C de três modos diferentes:

- $\text{BD1} = \text{Área A} + \text{Área B} + \text{Área C}$ , com 8 e 20 meses após a aplicação de gesso, contendo ao todo 106 registros.
- $\text{BD2} = \text{Área A} + \text{Área B}$ , com 8, 20, 32 e 44 meses após a aplicação de gesso, contendo ao todo 144 registros.
- $\text{BD3} = \text{Área B}$ , com 8, 20, 32, 44, 56, 68, 80 e 92 meses após a aplicação de gesso, contendo ao todo 132 registros.

### 3.4. Subdivisão das bases de dados

Antes de realizar a subdivisão das bases de dados BD1, BD2 e BD3 foi feita a análise dos dados contendo todos os atributos químicos coletados, além dos atributos época e gesso. Com o intuito de se trabalhar com menos atributos, antes mesmo de aplicar as técnicas de seleção de atributos, as bases de dados BD1, BD2 e BD3 foram subdivididas de acordo com atributos referentes ao teor e à saturação, e também à profundidade. Considerando que as amostras das camadas A (0–10 cm) e B (10–20 cm)

são superficiais e as das camadas C (20–40 cm) e D (40–60 cm) são do subsolo, calculou-se a média dos atributos das camadas A e B para obter os atributos da camada superficial de 0–20 cm (AB). Considerando também que tanto a camada C como a D são subsuperficiais, optou-se em trabalhar com os atributos da camada C separadamente como referência ao subsolo. Assim, para cada base de dados, BD1, BD2 e BD3, realizou-se a seguinte divisão:

- A = Todos os atributos químicos (teor + saturação) da camada superficial AB<sup>5</sup> e da camada subsuperficial C, mais os atributos época e gesso.
- B = Todos os atributos químicos (teor + saturação) de todas as camadas (A, B, C e D), mais os atributos época e gesso.
- C = Todos os atributos químicos (teor + saturação) da camada superficial AB, mais os atributos época e gesso.
- D = Todos os atributos químicos (teor + saturação) da camada subsuperficial C, mais os atributos época e gesso.
- E = Somente os atributos referentes ao teor: Al, Ca, Mg e K, da camada superficial AB, mais os atributos época e gesso.
- F = Somente os atributos referentes ao teor: Al, Ca, Mg e K, da camada subsuperficial C, mais os atributos época e gesso.
- G = Somente os atributos referentes à saturação: m, CaCTCe, MgCTCe e KCTCe, da camada superficial AB, mais os atributos época e gesso.
- H = Somente os atributos referentes à saturação: m, CaCTCe, MgCTCe e KCTCe, da camada subsuperficial C, mais os atributos época e gesso.
- I = Somente os atributos referentes ao teor: Al, Ca, Mg e K, de todas as camadas (A, B, C e D), mais os atributos época e gesso.
- J = Somente os atributos referentes à saturação: m, CaCTCe, MgCTCe e KCTCe, de todas as camadas (A, B, C e D), mais os atributos época e gesso.

### **3.5. Etapa de pré-processamento**

#### **3.5.1. Seleção de atributos**

O fato de ter sido calculada a média dos atributos das camadas A e B para obter os atributos da camada superficial de 0–20 cm (AB), de modo a diminuir a quantidade

---

<sup>5</sup> AB refere-se à média da camada A e camada B

de atributos, não exclui a necessidade de realizar a seleção de atributos, pois as bases de dados ainda continuam sendo de alta dimensionalidade.

Nesse trabalho empregaram-se os critérios de seleção B2 e B4, baseados em ACP, propostos por Jolliffe (1972). Utilizou-se também a técnica de ACPS proposta por Bair *et al.* (2006), com o critério B2. Para a aplicação da técnica de ACP empregando-se os critérios de seleção B2, B4 e ACPS foi empregada a ferramenta desenvolvida por Oliveira Jr. (2012).

Cramer (1999) propôs uma forma de estimar a matriz de covariância quando existe uma hipótese de independência marginal entre determinados atributos. Sejam  $L$ ,  $K$  e  $R$  três subconjuntos disjuntos de uma base de dados  $X$  tal que  $L \cup K \cup R \equiv X$  e  $L, K \neq \{\}$ . Se os atributos em  $L$  e  $K$  são independentes entre si, então a matriz de covariância  $\Omega_{K,L} = \text{cov}(x_K, x_L) = 0$ .

A estimativa da matriz de covariância assumindo relações de independência entre os atributos dos subconjuntos  $L$  e  $K$  é dada pela Equação 2 (CRAMER, 1999).

Equação 2 – Teorema proposto por Cramer (1999) assumindo independência entre  $K$  e  $L$

$$\Omega'' = \begin{bmatrix} \Omega' & \Omega' S'_{R*(L,K)} \\ S_{R*(L,K)} \Omega' & S_{R:(L,K)} + S_{R*(L,K)} \Omega S'_{R*(L,K)} \end{bmatrix}, \text{ em que } \Omega' = \begin{bmatrix} S_{L,L} & 0 \\ 0 & S_{K,K} \end{bmatrix}$$

Cramer (1999) introduziu duas notações matriciais que são utilizadas para o cálculo da matriz de covariância  $\Omega''$ . São elas:

- $S_{K*L} = S_{K,L} * S_{L,L}^{-1}$
- $S_{K:L} = S_{K,K} - S_{K,L} * S_{L,L}^{-1} * S'_{K,L}$ , onde  $S$  é a matriz de covariância amostral

A matriz de covariância  $\Omega''$  proposta por Cramer (1999) foi implementada na linguagem estatística R, utilizando a ferramenta R versão 2.14.0, e para realização da seleção dos atributos, depois de calculada a matriz de covariância  $\Omega''$  foram utilizados os critérios de seleção B2 e B4 baseadas em ACP. Para isso, a matriz de covariância  $\Omega''$  foi utilizada juntamente com a ferramenta de seleção desenvolvida por Oliveira Jr. (2012), em que ao invés de utilizar a matriz de covariância amostral  $S$  gerada pela ferramenta, empregou-se a matriz de covariância  $\Omega''$ . Uma possível independência marginal entre todos os atributos químicos do solo da camada A em relação aos atributos químicos da camada D foi avaliada por especialista da área. Assim, os testes envolvendo a matriz de covariância  $\Omega''$  foram realizados com base no conhecimento do especialista.

Pelo fato de ter sido utilizada indução de regras na realização desse trabalho, houve a necessidade de se trabalhar com poucos atributos, de modo que a regra fosse compreensível e válida. Então, em cada experimento, para cada critério de seleção de atributos, foram selecionados cinco, quatro e três atributos de cada base de dados. No entanto, também foram realizadas as análises com todos os atributos, antes da seleção dos mesmos.

### 3.5.2. Transformação das bases de dados

#### 3.5.2.1. SMOTE

O algoritmo SMOTE foi empregado para aumentar a quantidade de registros das bases de dados, utilizando-se do benefício do balanceamento dos dados. Esse algoritmo está implementado no software WEKA versão 3.6.5 (WITTEN e FRANK, 2005). O software WEKA (*Waikato Environment for Knowledge Analysis*) é amplamente utilizado por ter uma grande variedade de algoritmos implementados, desde técnicas de pré-processamento até algoritmos de tarefas de associação, classificação, regressão e agrupamento.

#### 3.5.2.2. Método mega-trend-diffusion

Outro método utilizado como técnica de ampliação de base de dados foi o método MTD proposto por Li *et al.* (2007a). O método foi implementado na linguagem de programação Java, utilizando a ferramenta Netbeans IDE versão 6.9.1.

O método MTD consiste em gerar amostras de dados virtuais, baseando-se nos dados já existentes. Para isso, o método cria o limite inferior e o limite superior do conjunto de amostra. Os limites são calculados a partir das seguintes fórmulas, sendo o limite inferior calculado através de  $a$  e o limite superior calculado através de  $b$ .

$$a = u_{set} - SkewL \sqrt{-2 * \frac{s_x^2}{NL * \ln(\varphi(a))}} \quad (1)$$

$$b = u_{set} + SkewU \sqrt{-2 * \frac{s_x^2}{NU * \ln(\varphi(b))}} \quad (2)$$

em que,  $u_{set}$  é calculado através da fórmula,  $u_{set} = \frac{(mínimo + máximo)}{2}$ .

Os valores de  $SkewL$  e  $SkewU$  são medidas de assimetria de distribuição dos dados e são calculados pelas fórmulas:

$$SkewL = \frac{NL}{(NL + NU)} \quad (3)$$

$$SkewU = \frac{NU}{(NU+NL)} \quad (4)$$

NL e NU encontrados nas Fórmulas (1), (2), (3) e (4) representam o número de dados que estão abaixo de  $u_{set}$  e o número de dados que estão acima  $u_{set}$ , respectivamente.

O valor de  $s_x^2$  é obtido através da variância do conjunto de amostra dada pela fórmula:

$$s_x^2 = \sum_{i=1}^n \frac{(xi-\bar{x})^2}{n-1} \quad (5)$$

em que,  $n$  é o tamanho da amostra. Finalmente, os valores de  $\varphi(a)$  e  $\varphi(b)$  são dados pelas seguintes fórmulas:

$$\varphi(a) = \varphi(r) = \varphi(0,9 + m * 10^{-2}) \quad (6)$$

$$\varphi(b) = \varphi(r) = \varphi(0,9 + m * 10^{-2}) \quad (7)$$

em que,  $r$  é o coeficiente de correlação linear dos dados de entrada em relação ao dado de saída, e  $m$  é o segundo dígito depois do ponto decimal de  $r$ . Assim, tem-se que  $\forall r \in \{0,91; 0,92; \dots; 0,99\}$ .

### 3.5.2.2.1. Exemplo numérico

Passo 1: Selecionar aleatoriamente uma parte da base de dados original. É permitido que o usuário defina quantos registros serão utilizados como conjunto de dados de treinamento. Então se separa os registros de acordo com os subconjuntos do atributo meta. Na Tabela 3 há quatro subconjuntos: 0, 4, 8 e 12.

Passo 2: Aplicar as Fórmulas (1) e (2) em cada subconjunto e calcular  $a$  e  $b$  para cada atributo. Para o subconjunto 12, há apenas um registro, então as Fórmulas (1) e (2) não podem ser aplicadas. Nessa situação aplicam-se as seguintes fórmulas para o cálculo de  $a$  e  $b$ .

$$a = \frac{\text{mínimo}}{5} \quad (8)$$

$$b = \frac{\text{máximo}}{5} \quad (9)$$

Outra situação que pode ocorrer é no atributo  $Ca_A$ , em que a variância do atributo é igual a zero. Nesse caso, também devem ser aplicadas as Fórmulas (8) e (9) ao invés das Fórmulas (1) e (2).

Passo 3: Após calculados os valores de  $a$  e  $b$  gera-se, aleatoriamente, as amostras artificiais dentro dos limites calculados. A quantidade de amostras artificiais que são criadas também é definida pelo usuário.

Passo 4: O processo é repetido até que toda a base de dados tenha sido amostrada.

Tabela 3 – Conjunto de dados de treinamento selecionadas aleatoriamente

| Ca <sub>A</sub> | Mg <sub>A</sub> | K <sub>A</sub> | Gesso |
|-----------------|-----------------|----------------|-------|
| 24              | 13,3            | 3,4            | 0     |
| 24              | 18              | 3,3            | 0     |
| 24              | 12              | 4,2            | 0     |
| 24,5            | 14              | 3,5            | 4     |
| 30              | 16              | 2,8            | 4     |
| 29,8            | 14              | 3,4            | 4     |
| 31              | 12              | 3              | 8     |
| 34,5            | 12,3            | 2,4            | 8     |
| 31              | 13,3            | 3,8            | 12    |

### 3.6. Etapa de processamento (MD)

Com base no objetivo de se chegar a um modelo para estimar a necessidade de gesso agrícola baseada nos atributos químicos do solo, utilizou-se o algoritmo de Aprendizagem de Máquina denominado *M5Rules*, o qual está implementado no software WEKA versão 6.8.1. Além da abordagem clássica da tarefa de regressão para a estimativa da necessidade do gesso agrícola, foi realizada também uma análise crítica sobre os modelos gerados, verificando sua coerência e utilidade com o auxílio do especialista humano na área de domínio. O algoritmo *M5Rules* foi aplicado em todas as subdivisões das bases de dados BD1, BD2 e BD3.

Os resultados obtidos foram analisados de acordo com o coeficiente de correlação ( $r$ ), o erro médio absoluto (EMA) e a importância da regra do ponto de vista agrônomo, de acordo com especialista da área.

### 3.7. Etapa de pós-processamento

O pós-processamento foi realizado com o auxílio de especialista da área para a atividade de visualização e interpretação dos resultados obtidos na etapa de MD, de modo que fossem analisadas apenas as regras coerentes agronomicamente.

Outra atividade realizada na etapa final do KDD foi a consolidação das regras, a qual foi realizada utilizando um subconjunto de registros, representando mais de 10% das bases de dados, sendo aproximadamente 14,2% da base BD1, 10,5% da base BD2 e 11,4% da base BD3 que continham as produtividades relativas mais altas (acima de 80%). A aplicação das regras para sua consolidação foi realizada da seguinte forma:

Dado um registro, representado na Tabela 4, a aplicação de uma regra foi feita pela substituição dos valores dos atributos que compõem a regra. Por exemplo, dada a regra:

$${}^6\text{gesso} = 0,075 \times \text{época} + 0,156 \times \text{CaCTCe}_{AB} - 0,302 \times \text{MgCTCe}_{AB} + 1,986,$$

a substituição dos valores dos atributos resulta em:

$$\text{gesso} = (0,075 \times 8) + (0,156 \times 69,35) - (0,302 \times 18,1) + 1,986 \cong 7,94$$

Tabela 4 – Representação de um registro para alguns atributos presentes na base de dados

| Local | Época | Gesso Apl. | Ca_AB | Mg_AB | Ca_CT Ce_AB | Mg_CT Ce_AB |
|-------|-------|------------|-------|-------|-------------|-------------|
| 1     | 8     | 4          | 41,0  | 8,0   | 69,35       | 18,1        |

### 3.8. Experimentos realizados

#### 3.8.1. Primeira fase

Na primeira fase dos experimentos trabalhou-se com as bases de dados BD1, BD2 e BD3 sem que houvesse a subdivisão das bases de dados quanto ao teor, saturação e profundidade. Antes de aplicar o algoritmo *M5Rules* nas três bases de dados, foi realizada a seleção de atributos seguindo o critério de seleção B2 baseada em ACP, de modo que foi realizada a seleção dos atributos permitindo a escolha dos cinco, quatro e três atributos mais relevantes da base de dados.

Após aplicação do algoritmo e com o auxílio de especialista da área, selecionaram-se as melhores regras do ponto de vista agrônomo. Com isso, aplicou-se a técnica SMOTE apenas nas bases de dados com as melhores regras.

Na Tabela 5 são mostradas as características dos experimentos realizados na primeira fase.

Tabela 5 – Experimentos da primeira fase

| Experimento | Base de dados | Subdivisão | Método de seleção | Critério de Seleção |
|-------------|---------------|------------|-------------------|---------------------|
| 1           | BD1           | A          | ACP               | B2                  |
| 2           | BD2           | A          | ACP               | B2                  |
| 3           | BD3           | A          | ACP               | B2                  |

#### 3.8.2. Segunda fase

Na segunda fase dos experimentos trabalhou-se com as subdivisões C, D, E, F, G e H, das bases BD1, BD2 e BD3. Na etapa de pré-processamento dessa fase foram utilizados os critérios de seleção de atributos B2, B4 baseada em ACP e o método

<sup>6</sup> Regra definida apenas para a demonstração da atividade de consolidação

ACPS com o critério B2. A seleção de atributos foi realizada selecionando apenas os cinco, quatro e três atributos mais relevantes da base. Ainda na etapa de pré-processamento, foram aplicados os métodos MTD e SMOTE para ampliação da base de dados. Por fim, na etapa de MD foi aplicado o algoritmo *M5Rules* em todas as subdivisões das bases de dados. No entanto, o método MTD foi aplicado em uma segunda parte do trabalho que objetivava apenas a análise dos resultados computacionais (*r* e EMA).

Na Tabela 6 são mostradas as características dos experimentos realizados na segunda fase.

Tabela 6 – Experimentos da segunda fase

| Experimento | Base de dados | Subdivisão | Método de Seleção | Critério de Seleção |
|-------------|---------------|------------|-------------------|---------------------|
| 4           | BD1           | C          | ACP e ACPS        | B2 e B4             |
| 5           | BD1           | D          | ACP e ACPS        | B2 e B4             |
| 6           | BD1           | E          | ACP e ACPS        | B2 e B4             |
| 7           | BD1           | F          | ACP e ACPS        | B2 e B4             |
| 8           | BD1           | G          | ACP e ACPS        | B2 e B4             |
| 9           | BD1           | H          | ACP e ACPS        | B2 e B4             |
| 10          | BD2           | C          | ACP e ACPS        | B2 e B4             |
| 11          | BD2           | D          | ACP e ACPS        | B2 e B4             |
| 12          | BD2           | E          | ACP e ACPS        | B2 e B4             |
| 13          | BD2           | F          | ACP e ACPS        | B2 e B4             |
| 14          | BD2           | G          | ACP e ACPS        | B2 e B4             |
| 15          | BD2           | H          | ACP e ACPS        | B2 e B4             |
| 16          | BD3           | C          | ACP e ACPS        | B2 e B4             |
| 17          | BD3           | D          | ACP e ACPS        | B2 e B4             |
| 18          | BD3           | E          | ACP e ACPS        | B2 e B4             |
| 19          | BD3           | F          | ACP e ACPS        | B2 e B4             |
| 20          | BD3           | G          | ACP e ACPS        | B2 e B4             |
| 21          | BD3           | H          | ACP e ACPS        | B2 e B4             |

### 3.8.2. Terceira fase

A terceira fase dos experimentos consistiu em utilizar o conhecimento *a priori* sobre asserções de independência marginal entre os atributos químicos das camadas A e D. Para isso foram realizados os cálculos da matriz de covariância  $\Omega''$  proposta por Cramer (1999).

Para efeito comparativo foi também realizado o mesmo procedimento em todos os casos utilizando a matriz de covariância amostral *S* calculada pela própria ferramenta de Oliveira Jr. (2012). Nos dois casos, a seleção de atributos foi realizada selecionando apenas os cinco, quatro e três atributos mais relevantes das bases de dados.

Conforme as duas primeiras fases, na etapa de MD, foi utilizado o algoritmo *M5Rules*. Na Tabela 7 podem ser observadas as características dos experimentos realizados na terceira fase, sendo que para todos os experimentos o método de seleção adotado foi o ACP e os critérios de seleção utilizados também em todos os experimentos foram o B2 e B4.

Tabela 7 – Experimentos da segunda fase

| <b>Experimentos</b> | <b>Base de dados</b> | <b>Subdivisão</b> | <b>Matriz de covariância</b> |
|---------------------|----------------------|-------------------|------------------------------|
| <b>22</b>           | BD1                  | B                 | $\Omega''$                   |
| <b>23</b>           | BD1                  | I                 | $\Omega''$                   |
| <b>24</b>           | BD1                  | J                 | $\Omega''$                   |
| <b>25</b>           | BD1                  | B                 | S                            |
| <b>26</b>           | BD1                  | I                 | S                            |
| <b>27</b>           | BD1                  | J                 | S                            |
| <b>28</b>           | BD2                  | B                 | $\Omega''$                   |
| <b>29</b>           | BD2                  | I                 | $\Omega''$                   |
| <b>30</b>           | BD2                  | J                 | $\Omega''$                   |
| <b>31</b>           | BD2                  | B                 | S                            |
| <b>32</b>           | BD2                  | I                 | S                            |
| <b>33</b>           | BD2                  | J                 | S                            |
| <b>34</b>           | BD3                  | B                 | $\Omega''$                   |
| <b>35</b>           | BD3                  | I                 | $\Omega''$                   |
| <b>36</b>           | BD3                  | J                 | $\Omega''$                   |
| <b>37</b>           | BD3                  | B                 | S                            |
| <b>38</b>           | BD3                  | I                 | S                            |
| <b>39</b>           | BD3                  | J                 | S                            |

## 4. RESULTADOS E DISCUSSÃO

### 4.1. Estimativa de gesso agrícola

#### 4.1.1. Primeira fase

Os experimentos realizados durante a primeira fase não apresentaram resultados satisfatórios para a estimativa da necessidade de gesso agrícola. Ao utilizar as bases de dados BD1, BD2 e BD3 com a subdivisão A, foram geradas algumas regras com  $r$  satisfatórios (acima de 0,75), porém incoerentes do ponto de vista agrônomo. Como exemplo, pode-se citar a Regra (1) obtida a partir da base BD1 considerando todos os atributos, cujo valor de  $r = 0,84$  e  $EMA = 1,65$ .

$$(1) \quad gesso = 0,0959 \times \acute{e}poca + 1,0021 \times Al_{AB} - 0,2187 \times Ca_C + 0,1141 \times Mg_{AB} + 0,2193 \times Mg_C + 1,7824 \times K_C + 0,3183 \times CaCTCe_{AB} + 0,2188 \times CaCTCe_C - 0,5856 \times KCTCe_C - 29,3984$$

Na Regra (1), existem atributos referentes tanto ao teor quanto à saturação de cátions, o que torna a regra inviável agronomicamente. Além disso, essa regra é muito específica por conter vários atributos, o que confere baixa cobertura da base de dados.

Ao ser aplicado o critério B2 na seleção de atributos baseada em ACP, buscando-se regras contendo apenas os cinco, quatro e três atributos mais relevantes das bases de dados BD1, BD2 e BD3, as regras geradas, ainda assim, não foram satisfatórias pelo fato de conter atributos de teor e saturação no mesmo modelo.

Na Regra (1), a relação inversamente proporcional do atributo  $Ca_C$  em relação ao gesso está incoerente com a realidade agrônoma, uma vez que o gesso é excelente fonte Ca para o solo (OLIVEIRA e PAVAN, 1996; CAIRES *et al.*, 2001; MARQUES, 2008; NEIS *et al.*, 2010). Como a saturação é obtida a partir do teor de cátions, se ambos estão presentes na regra eles devem ter a mesma relação com o atributo que está sendo estimado. Isso não acontece com os atributos  $Ca_C$  e  $CaCTCe_C$ , sendo o primeiro com relação inversa e o segundo com relação diretamente proporcional.

A compreensão de que a Regra (1) não é satisfatória se fez justamente pelo fato desse estudo ter sido conduzido com o acompanhamento constante de especialista da área, o que foi indispensável para a determinação da utilidade e coerência dos resultados obtidos em todas as fases desse trabalho.

As regras geradas após a aplicação do método SMOTE mantiveram as mesmas características anteriores. Embora a aplicação do método SMOTE tenha resultado em pequeno aumento do valor do  $r$  e redução do valor do EMA, do ponto de vista do domínio da aplicação as regras se mostraram inviáveis.

#### 4.1.2. Segunda fase

Em oposto à primeira, na segunda fase os resultados foram interessantes tanto do ponto vista agrônomo, para a estimativa da necessidade de gesso agrícola, quanto do ponto vista computacional, com valores aceitáveis de  $r$  e EMA. Tal fato se justifica pelas subdivisões C, D, E, F, G e H nas bases BD1, BD2 e BD3.

Ao analisar a base de dados contendo registros de todas as áreas (BD1) considerando todos os atributos referentes ao teor e saturação na superfície, além dos atributos época e gesso (subdivisão C), foram geradas algumas regras interessantes, como por exemplo:

$$(2) \quad \text{gesso} = 0,0888 \times \text{época} + 0,1482 \times \text{CaCTCe}_{AB} - 0,2998 \times \text{MgCTCe}_{AB} + 2,3871$$

Essa regra apresentou  $r = 0,82$  e  $\text{EMA} = 1,77$ , e foi gerada a partir do subconjunto dos atributos época,  $\text{Ca}_{AB}$ ,  $\text{CaCTCe}_{AB}$  e  $\text{MgCTCe}_{AB}$ , que foram selecionados pelo critério B2 baseado em ACP.

Apesar de essa subdivisão englobar tanto os atributos referentes ao teor quanto à saturação, os critérios B2, B4 e o método ACPS selecionaram, em sua maioria, os atributos envolvendo a saturação como os que melhor representaram a variação dos dados.

Apesar de a seleção utilizando o critério B2 para a Regra (2) incluir o atributo  $\text{Ca}_{AB}$ , o mesmo não compôs a regra, pois além da seleção de atributos utilizada na etapa de pré-processamento, o algoritmo *M5Rules* também realiza seleção de atributos que leva em consideração o ganho de informação do atributo e, baseado nesse critério, esse algoritmo decide por incluir ou não o atributo na regra. Com isso, percebe-se que houve maior importância da saturação em relação aos atributos referentes ao teor.

Os atributos envolvendo a saturação explicam melhor o comportamento dos dados quando a análise dos mesmos envolve solos com características e concentrações de nutrientes diferentes, pois ao trabalhar com a saturação ocorre uniformização dos dados, de modo que os valores de saturação passam a ser válidos para os diferentes solos (Raij, 1991). É este o caso das bases BD1, que envolve as áreas de estudo A, B e C, e a base BD2 que envolve as áreas A e B. Isso explica o motivo pelo qual apenas os atributos envolvendo a saturação compuseram a Regra (2), uma vez que a mesma foi obtida a partir dos dados sobre a base BD1.

Ao utilizar o método B4, selecionando apenas os três atributos mais relevantes, resultou-se nos atributos época,  $\text{CaCTCe}_{AB}$  e  $\text{MgCTCe}_{AB}$ . A regra gerada nessas

condições foi idêntica à Regra (2), e com  $r$  e EMA semelhantes, sendo, 0,83 e 1,75, respectivamente.

O mesmo aconteceu ao utilizar o método de seleção ACPS, em que selecionando os cinco atributos mais relevantes da base BD1 com subdivisão C obtiveram-se os seguintes atributos: época,  $Ca_{AB}$ ,  $Mg_{AB}$ ,  $CaCTCe_{AB}$  e  $MgCTCe_{AB}$ . Com isso, foi gerada uma regra igual à Regra (2), com  $r = 0,81$  e  $EMA = 1,78$ .

A pouca diferença de  $r$  e EMA deu-se pela validação da regra por meio do conjunto de testes que foi selecionado aleatoriamente. Com isso, a aplicação da Regra (2) foi testada em diferentes conjuntos de teste, uma vez que foi gerada três vezes.

Vale ressaltar a importância dos atributos envolvendo a saturação por Ca e saturação por Mg da camada superficial do solo (0–20 cm), pois nos três critérios de seleção, ambos atributos foram selecionados como dois dos três atributos mais relevantes da base de dados. Além disso, os valores de saturação são obtidos pelos teores dos mesmos e, com isso, é coerente a presença dos teores de Ca e Mg no subconjunto de atributos selecionados pelo critério B2 e pelo método ACPS.

A Regra (2) apresenta coerência com os trabalhos de Oliveira e Pavan (1996), Caires et al. (1998), Caires et al. (2004) e Soratto e Cruciol (2008a), em que a aplicação do gesso aumentou os teores de Ca e diminuiu os teores de Mg nas camadas superficiais do solo. Isso pode ser observado por meio do comportamento dos atributos em relação à dose gesso estimada, em que o  $CaCTCe_{AB}$  teve correlação diretamente proporcional (+0,1482) enquanto que o  $MgCTCe_{AB}$  apresentou correlação inversamente proporcional (−0,2998), levando ao entendimento de que a medida em que a dose de gesso aumentou, o  $CaCTCe_{AB}$  se elevou ao mesmo tempo em que o  $MgCTCe_{AB}$  diminuiu.

As estimativas do gesso geradas pela aplicação da Regra (2) podem ser visualizadas no Quadro 1, que representam aproximadamente 14,2% da base de dados.

Quadro 1 – Estimativa para aplicação de gesso utilizando a Regra (2)

| LINHA | ESTIMATIVA<br>(t ha <sup>-1</sup> ) | QTDE. APLICADA<br>(t ha <sup>-1</sup> ) | DIFERENÇA<br>(t ha <sup>-1</sup> ) | ÉPOCA<br>(meses após aplicação) |
|-------|-------------------------------------|-----------------------------------------|------------------------------------|---------------------------------|
| 1     | 8,0                                 | 4                                       | 4,0                                | 8                               |
| 2     | 8,8                                 | 8                                       | 0,8                                | 20                              |
| 3     | 8,8                                 | 9                                       | 0,2                                | 8                               |
| 4     | 6,6                                 | 6                                       | 0,6                                | 20                              |
| 5     | 11,2                                | 6                                       | 5,2                                | 32                              |
| 6     | 3,9                                 | 3                                       | 0,9                                | 44                              |
| 7     | 7,9                                 | 9                                       | 1,1                                | 56                              |

|    |      |   |     |    |
|----|------|---|-----|----|
| 8  | 11,1 | 6 | 5,1 | 80 |
| 9  | 15,0 | 6 | 9,0 | 92 |
| 10 | 8,4  | 8 | 0,4 | 8  |
| 11 | 2,4  | 4 | 1,6 | 20 |
| 12 | 2,5  | 4 | 1,5 | 20 |
| 13 | 4,3  | 6 | 1,7 | 44 |
| 14 | 13,8 | 9 | 4,8 | 80 |
| 15 | 7,7  | 8 | 0,3 | 8  |
| 16 | 5,5  | 4 | 1,5 | 8  |

Apesar de a Regra (2) conter, em alguns casos, diferenças relativamente grandes entre a dose estimativa e quantidade real aplicada de gesso, como no caso da linha 9, em que a diferença foi de 9 t ha<sup>-1</sup>, na maioria dos casos a diferença pode ser considerada aceitável. Em onze casos encontrados no Quadro 1, a diferença foi igual ou menor que 1,7 t ha<sup>-1</sup>.

Outra regra que foi gerada mais de uma vez foi a Regra (3), a qual foi gerada, primeiramente, a partir do subconjunto de atributos Ca<sub>AB</sub>, CaCTCe<sub>AB</sub>, MgCTCe<sub>AB</sub> selecionados pelo critério B2. O método ACPS selecionou o mesmo subconjunto de atributos e, com isso, gerou-se a mesma regra.

$$(3) \text{ gesso} = 0,1675 \times \text{CaCTCe}_{AB} - 0,2647 \times \text{MgCTCe}_{AB} + 1,4675$$

Como o subconjunto de atributos foi o mesmo, tanto selecionado pelo critério B2 quanto pelo método ACPS, os valores de  $r$  (0,81) e EMA (1,74) foram os mesmos e, além disso, foram considerados satisfatórios. A Regra (3) foi composta pelos mesmos atributos da Regra (2), exceto para o atributo época que, nesse caso, não se apresentou como um dos três atributos mais relevantes nos critério B2 e no método ACPS.

Na base BD2, ainda com a subdivisão C, houve apenas uma regra merecendo destaque, sendo:

$$(4) \text{ gesso} = 0,0664 \times \text{época} + 0,2748 \times \text{CaCTCe}_{AB} - 12,714$$

A Regra (4) apresentou  $r = 0,63$  e EMA = 2,42, e foi gerada a partir dos dados dos atributos época, Ca<sub>AB</sub> e CaCTCe<sub>AB</sub>, selecionados no pré-processamento pelo método ACPS. As demais regras geradas ficaram confusas pelo mesmo motivo que ocorreu na primeira fase, contendo, ao mesmo tempo, atributos relativos tanto ao teor como à saturação.

O Quadro 2 representa a consolidação da Regra (4). É possível notar que houve oito casos em que a diferença entre a dose estimada e a quantidade real aplicada de gesso foi menor que 1 t ha<sup>-1</sup>.

Quadro 2 – Estimativa para aplicação de gesso utilizando a Regra (4)

| LINHA | ESTIMATIVA<br>(t ha <sup>-1</sup> ) | QTDE. APLICADA<br>(t ha <sup>-1</sup> ) | DIFERENÇA<br>(t ha <sup>-1</sup> ) | ÉPOCA<br>(meses após aplicação) |
|-------|-------------------------------------|-----------------------------------------|------------------------------------|---------------------------------|
| 1     | 6,9                                 | 4                                       | 2,9                                | 8                               |
| 2     | 8,7                                 | 8                                       | 0,7                                | 20                              |
| 3     | 8,2                                 | 9                                       | 0,8                                | 8                               |
| 4     | 7,0                                 | 6                                       | 1,0                                | 20                              |
| 5     | 8,5                                 | 6                                       | 2,5                                | 32                              |
| 6     | 4,5                                 | 3                                       | 1,5                                | 44                              |
| 7     | 8,1                                 | 9                                       | 0,9                                | 56                              |
| 8     | 6,4                                 | 6                                       | 0,4                                | 80                              |
| 9     | 9,8                                 | 6                                       | 3,8                                | 92                              |
| 10    | 8,9                                 | 8                                       | 0,9                                | 8                               |
| 11    | 5,2                                 | 4                                       | 1,2                                | 20                              |
| 12    | 3,5                                 | 4                                       | 0,5                                | 20                              |
| 13    | 5,3                                 | 6                                       | 0,7                                | 44                              |
| 14    | 11,0                                | 9                                       | 2,0                                | 80                              |
| 15    | 8,6                                 | 8                                       | 0,6                                | 8                               |
| 16    | 7,1                                 | 4                                       | 3,1                                | 8                               |

Já, na base BD3 com subdivisão C, não foram encontradas boas regras do ponto de vista agrônomo para a estimativa da necessidade de gesso, apesar de, em alguns casos, ter sido encontradas regras com valores de  $r$  em torno de 0,75. Isso reforça ainda mais a importância de um especialista na etapa de pós-processamento do KDD, em que é realizada a análise e interpretação dos resultados, uma vez que a regra, além de correta, deve ser útil e interessante.

Para a subdivisão D, que utilizou os atributos envolvendo tanto o teor quanto a saturação, além dos atributos época e gesso, da camada subsuperficial, as bases BD1, BD2 e BD3 não apresentaram nenhuma regra considerada aplicável para a estimativa de gesso. Além disso, numericamente, as regras apresentaram baixo valor de  $r$ , como por exemplo, a Regra (5) gerada a partir da base BD3. Essa regra apresentou  $r = 0,42$  e  $EMA = 2,56$  e, nesse caso, os atributos foram selecionados pelo método ACPS, sendo eles:  $Al_C$ ,  $Ca_C$ ,  $Mg_C$ ,  $CaCTCe_C$  e  $MgCTCe_C$ .

$$(5) \text{ gesso} = 0,179 \times Ca_C + 0,2375 \times Mg_C - 0,1599 \times MgCTCe_C + 4,257$$

O algoritmo não conseguiu gerar estimativa satisfatória, resultando em regras contraditórias, como no caso da Regra (5) em que o teor de Mg esteve diretamente relacionado com a aplicação de gesso, em desacordo com os trabalhos de Quaggio *et al.* (1993), Caires *et al.* (2004, 2011). A relação inversamente proporcional da saturação

por Mg com o gesso, apesar de estar correta, está incoerente com a relação do gesso e teor de Mg, uma vez que era esperado que ambos os atributos apresentassem a mesma proporcionalidade na correlação.

As Tabelas 8, 9 e 10 apresentam os valores de  $r$  e EMA de todas as análises realizadas com o algoritmo *M5Rules* sobre a base BD1, BD2 e BD3, respectivamente, com as subdivisões C e D. Nas três tabelas, os resultados de  $r$  e EMA foram, em sua maioria, melhores quando foi utilizada a subdivisão C, sendo este resultado mais evidente na Tabela 8. Com isso, pode-se afirmar que, na maioria dos casos, os atributos envolvendo a camada superficial explicaram melhor o comportamento dos dados do que aqueles da camada subsuperficial.

Os valores de  $r$  e EMA presentes nas Tabelas 8, 9 e 10 são considerados como  $r$  e EMA médios da execução do algoritmo, sendo que em cada execução podem ter sido geradas  $n$  regras por meio do *M5Rules*, embora, na maioria dos casos, o algoritmo gerou apenas uma regra por execução.

Tabela 8 – Resultados do  $r$  e EMA na base BD1 com subdivisão C e D

| Critério de SA <sup>7</sup> | Qtde. de AR <sup>8</sup> | BD1          |      |              |      |
|-----------------------------|--------------------------|--------------|------|--------------|------|
|                             |                          | Subdivisão C |      | Subdivisão D |      |
|                             |                          | $r$          | EMA  | $R$          | EMA  |
| Nenhum                      | 9                        | 0,85         | 1,63 | 0,80         | 1,95 |
| B2                          | 5                        | 0,81         | 1,78 | 0,77         | 2,05 |
| B2                          | 4                        | 0,82         | 1,77 | 0,76         | 2,01 |
| B2                          | 3                        | 0,81         | 1,75 | 0,76         | 2,00 |
| B4                          | 5                        | 0,83         | 1,79 | 0,78         | 1,96 |
| B4                          | 4                        | 0,86         | 1,58 | 0,76         | 2,01 |
| B4                          | 3                        | 0,83         | 1,75 | -0,10        | 3,69 |
| ACPS                        | 5                        | 0,81         | 1,78 | 0,77         | 1,93 |
| ACPS                        | 4                        | 0,82         | 1,77 | 0,74         | 2,13 |
| ACPS                        | 3                        | 0,81         | 1,75 | 0,78         | 1,92 |
| <b>Média</b>                |                          | 0,82         | 1,74 | 0,68         | 2,16 |
| <b>Desvio Padrão</b>        |                          | 0,02         | 0,07 | 0,27         | 0,54 |

<sup>7</sup> SA: Seleção de Atributos

<sup>8</sup> AR: Atributos Relevantes

Tabela 9 – Resultados do  $r$  e EMA na base BD2 com subdivisão C e D

| Critério de SA       | Qntd. De AR | BD2          |      |              |      |
|----------------------|-------------|--------------|------|--------------|------|
|                      |             | Subdivisão C |      | Subdivisão D |      |
|                      |             | $r$          | EMA  | $R$          | EMA  |
| Nenhum               | 9           | 0,73         | 2,07 | 0,68         | 2,30 |
| B2                   | 5           | 0,74         | 2,05 | 0,70         | 2,24 |
| B2                   | 4           | 0,61         | 2,42 | 0,65         | 2,48 |
| B2                   | 3           | 0,59         | 2,45 | 0,61         | 2,56 |
| B4                   | 5           | 0,66         | 2,35 | 0,56         | 2,69 |
| B4                   | 4           | 0,63         | 2,45 | 0,48         | 2,88 |
| B4                   | 3           | 0,59         | 2,53 | 0,24         | 3,42 |
| ACPS                 | 5           | 0,74         | 2,08 | 0,68         | 2,29 |
| ACPS                 | 4           | 0,74         | 2,05 | 0,65         | 2,44 |
| ACPS                 | 3           | 0,63         | 2,42 | 0,64         | 2,40 |
| <b>Média</b>         |             | 0,67         | 2,29 | 0,59         | 2,57 |
| <b>Desvio Padrão</b> |             | 0,06         | 0,20 | 0,14         | 0,36 |

Tabela 10 – Resultados do  $r$  e EMA na base BD3 com subdivisão C e D

| Critério de SA       | Qntd. de AR | BD3          |      |              |      |
|----------------------|-------------|--------------|------|--------------|------|
|                      |             | Subdivisão C |      | Subdivisão D |      |
|                      |             | $r$          | EMA  | $R$          | EMA  |
| Nenhum               | 9           | 0,74         | 1,68 | 0,49         | 2,41 |
| B2                   | 5           | 0,75         | 1,69 | 0,51         | 2,44 |
| B2                   | 4           | 0,72         | 1,81 | 0,47         | 2,52 |
| B2                   | 3           | 0,05         | 2,97 | 0,49         | 2,46 |
| B4                   | 5           | 0,74         | 1,72 | 0,44         | 2,54 |
| B4                   | 4           | 0,35         | 2,66 | 0,44         | 2,49 |
| B4                   | 3           | 0,05         | 2,97 | 0,44         | 2,53 |
| ACPS                 | 5           | 0,73         | 1,79 | 0,42         | 2,45 |
| ACPS                 | 4           | 0,72         | 1,82 | 0,45         | 2,52 |
| ACPS                 | 3           | 0,65         | 2,04 | 0,47         | 2,46 |
| <b>Média</b>         |             | 0,55         | 2,12 | 0,46         | 2,48 |
| <b>Desvio Padrão</b> |             | 0,29         | 0,54 | 0,03         | 0,04 |

Como as bases de dados BD1 e BD2 contêm diferentes tipos de solo e, por isso, as concentrações também são diferentes, isso dificultou as análises envolvendo apenas o teor para essas bases. Com isso, para as subdivisões E e F, em que foram considerados somente os atributos referentes ao teor da superfície (E) e subsuperfície (F), apenas as regras geradas a partir da base BD3 foram levadas em consideração, por ser a única contendo dados somente de uma área de estudo, no caso a área B.

Na subdivisão E, utilizando o critério de seleção de atributos B2, com os três atributos mais relevantes: época,  $Ca_{AB}$  e  $Mg_{AB}$ , obteve-se a seguinte regra:

$$(6) \text{ gesso} = 0,0396 \times \text{epoca} + 0,2891 \times Ca_{AB} - 0,2809 \times Mg_{AB} - 1,8504$$

A Regra (6), além de ter apresentado valores de  $r$  (0,71) e EMA (1,86) considerados satisfatórios, também foi viável agronomicamente, estando de acordo com os trabalhos de Oliveira e Pavan (1996), Caires *et al.*, (1998, 2004), e Soratto e Cruciol (2008a), nos quais a aplicação de gesso aumentou o teor de Ca e diminuiu o teor de Mg na camada superficial do solo (0–20 cm).

Essa mesma regra foi gerada também quando foram selecionados, com o uso do critério B4, os atributos época,  $Ca_{AB}$  e  $Mg_{AB}$ , como os três atributos mais relevantes da base de dados. Por serem exatamente os mesmos atributos selecionados, os valores de  $r$  e EMA foram iguais aos apresentados na Regra (6).

A consolidação da Regra (6) foi realizada utilizando um subconjunto de dados da base BD3, obtendo-se os resultados constantes no Quadro 3. Na maioria dos casos, como nas linhas 2, 4, 5, 6, 10, 12, 13, 15 e 16, houve boa estimativa da dose de gesso, sendo a diferença menor que  $1 \text{ t ha}^{-1}$ .

Quadro 3 – Estimativa para aplicação de gesso utilizando a Regra (6)

| LINHA | ESTIMATIVA<br>(t ha <sup>-1</sup> ) | QTDE. APLICADA<br>(t ha <sup>-1</sup> ) | DIFERENÇA<br>(t ha <sup>-1</sup> ) | ÉPOCA<br>(meses após aplicação) |
|-------|-------------------------------------|-----------------------------------------|------------------------------------|---------------------------------|
| 1     | 10,8                                | 9                                       | 1,8                                | 8                               |
| 2     | 9,2                                 | 9                                       | 0,2                                | 8                               |
| 3     | 7,4                                 | 6                                       | 1,4                                | 20                              |
| 4     | 3,0                                 | 3                                       | 0,0                                | 20                              |
| 5     | 6,0                                 | 6                                       | 0,0                                | 32                              |
| 6     | 2,8                                 | 3                                       | 0,2                                | 44                              |
| 7     | 6,2                                 | 9                                       | 2,8                                | 56                              |
| 8     | 4,9                                 | 6                                       | 1,1                                | 80                              |
| 9     | 5,2                                 | 3                                       | 2,2                                | 80                              |
| 10    | 6,9                                 | 6                                       | 0,9                                | 92                              |
| 11    | 4,0                                 | 6                                       | 2,0                                | 8                               |
| 12    | 6,6                                 | 6                                       | 0,6                                | 20                              |
| 13    | 9,4                                 | 9                                       | 0,4                                | 20                              |
| 14    | 3,7                                 | 6                                       | 2,3                                | 44                              |
| 15    | 9,9                                 | 9                                       | 0,9                                | 92                              |
| 16    | 3,2                                 | 3                                       | 0,2                                | 56                              |

Ainda na subdivisão E, outra regra obtida, nas mesmas condições, foi a Regra (7), que apresentou  $r = 0,72$  e  $EMA = 1,80$ , gerada a partir dos atributos época,  $Ca_{AB}$ ,  $Mg_{AB}$  e  $K_{AB}$ , que foram selecionados pelo critério de seleção B4.

$$(7) \text{ gesso} = 0,032 \times \text{época} + 0,2842 \times Ca_{AB} - 0,2613 \times Mg_{AB} - 0,5218 \times K_{AB} + 0,1077$$

O Quadro 4 representa a aplicabilidade da Regra (7) sobre o subconjunto de registros para testes. Dentre os 18 registros testados, apenas nas linhas 1 e 8 foram obtidas diferenças acima de  $2 \text{ t ha}^{-1}$  entre a dose estimada e a quantidade aplicada de gesso, sendo que onze casos apresentaram diferenças menores que  $1 \text{ t ha}^{-1}$ .

Quadro 4 – Estimativa para aplicação de gesso utilizando a Regra (7)

| LINHA | ESTIMATIVA<br>(t ha <sup>-1</sup> ) | QTDE. APLICADA<br>(t ha <sup>-1</sup> ) | DIFERENÇA<br>(t ha <sup>-1</sup> ) | ÉPOCA<br>(meses após aplicação) |
|-------|-------------------------------------|-----------------------------------------|------------------------------------|---------------------------------|
| 1     | 11,1                                | 9                                       | 2,1                                | 8                               |
| 2     | 9,3                                 | 9                                       | 0,3                                | 8                               |
| 3     | 6,7                                 | 6                                       | 0,7                                | 20                              |
| 4     | 2,3                                 | 3                                       | 0,7                                | 20                              |
| 5     | 6,1                                 | 6                                       | 0,1                                | 32                              |
| 6     | 3,4                                 | 3                                       | 0,4                                | 44                              |
| 7     | 5,3                                 | 6                                       | 0,7                                | 80                              |
| 8     | 5,9                                 | 3                                       | 2,9                                | 80                              |
| 9     | 6,2                                 | 6                                       | 0,2                                | 92                              |
| 10    | 4,1                                 | 6                                       | 1,9                                | 8                               |
| 11    | 6,4                                 | 6                                       | 0,4                                | 20                              |
| 12    | 9,3                                 | 9                                       | 0,3                                | 20                              |
| 13    | 4,4                                 | 6                                       | 1,6                                | 44                              |
| 14    | 8,9                                 | 9                                       | 0,1                                | 92                              |
| 15    | 3,0                                 | 3                                       | 0,0                                | 56                              |
| 16    | 4,9                                 | 3                                       | 1,9                                | 20                              |

Apesar de a lixiviação de K não ser encontrada em todos os trabalhos envolvendo o uso de gesso agrícola, esse é um efeito esperado, conforme observado nos trabalhos de Souza e Ritchey (1986) e Caires *et al.* (1998). Com isso, a Regra (7) também está coerente com a literatura porque na camada superficial do solo (0–20 cm), além do teor de Mg, o atributo  $K_{AB}$  também apresentou correlação negativa com o atributo gesso.

Na subdivisão F, em que foram utilizados apenas os atributos referentes aos teores da subsuperfície, mais os atributos época e gesso, nenhuma das regras geradas foi considerada viável para a estimativa do gesso, tendo apresentado baixos valores de  $r$  e

EMA, como pode ser visto na Tabela 11. Esses resultados reforçam a idéia de que os atributos da camada superficial tenderam a gerar regras melhores para a estimativa da necessidade de gesso. Assim, como nas Tabelas 8, 9 e 10, os valores de  $r$  e EMA presentes na Tabela 11 representam, respectivamente, o  $r$  e EMA médios obtidos.

Tabela 11 – Resultados do  $r$  e EMA na base BD3 com subdivisão E e F

| Critério de SA       | Qntd. De AR | BD3          |      |              |      |
|----------------------|-------------|--------------|------|--------------|------|
|                      |             | Subdivisão E |      | Subdivisão F |      |
|                      |             | $r$          | EMA  | $r$          | EMA  |
| Nenhum               | 9           | 0,75         | 1,69 | 0,43         | 2,59 |
| B2                   | 4           | 0,75         | 1,71 | 0,48         | 2,50 |
| B2                   | 3           | 0,71         | 1,86 | 0,47         | 2,52 |
| B4                   | 4           | 0,72         | 1,80 | 0,49         | 2,47 |
| B4                   | 3           | 0,71         | 1,86 | 0,47         | 2,52 |
| ACPS                 | 4           | 0,73         | 1,77 | 0,49         | 2,47 |
| ACPS                 | 3           | 0,68         | 1,97 | 0,48         | 2,49 |
| <b>Média</b>         |             | 0,72         | 1,81 | 0,47         | 2,51 |
| <b>Desvio Padrão</b> |             | 0,02         | 0,10 | 0,02         | 0,04 |

Ao contrário das subdivisões E e F, para as subdivisões G e H, envolvendo saturação da superfície e subsuperfície, respectivamente, levou-se em consideração todas as bases de dados (BD1, BD2 e BD3), pois estas subdivisões envolveram somente os atributos de saturação. Na base BD1, com subdivisão G, houve regras com valores de  $r$  considerados muito bons, como é o caso da Regra (8).

$$(8) \quad \text{gesso} = 0,1168 \times \text{época} + 0,0051 \times \text{CaCTCe}_{AB} - 0,3459 \times \text{MgCTCe}_{AB} - 0,3584 \times \text{KCTCe}_{AB} + 9,7584$$

Esta regra apresentou valores de  $r$  (0,84) e EMA (1,70) satisfatórios, e foi gerada a partir dos atributos época, CaCTCe<sub>AB</sub>, MgCTCe<sub>AB</sub> e KCTCe<sub>AB</sub> que foram selecionados pelo método ACPS. O Quadro 5 representa a consolidação da Regra (8).

Para essa regra, houve grandes diferenças entre a dose estimada e a quantidade real de gesso aplicado, como é o caso das linhas 1, 5, 8, 9 e 14, em que a estimativa foi superior a 3,9 t ha<sup>-1</sup>, chegando a alcançar uma diferença de até 10,5 t ha<sup>-1</sup>. Uma possível explicação foi a presença do atributo KCTCe<sub>AB</sub>, pois a lixiviação do K não ocorreu em todas as áreas de estudo. Como a base BD1 envolve as três áreas de estudo (A, B e C), é importante ressaltar que nas Áreas A e C constatou-se lixiviação do K (CAIRES *et al.*, 1998, 2011), mas na Área B os teores de K trocável no solo não foram influenciados pela aplicação de gesso (CAIRES *et al.*, 2004).

Quadro 5 – Estimativa para aplicação de gesso utilizando a Regra (8)

| LINHA | ESTIMATIVA<br>(t ha <sup>-1</sup> ) | QTDE. APLICADA<br>(t ha <sup>-1</sup> ) | DIFERENÇA<br>(t ha <sup>-1</sup> ) | ÉPOCA<br>(meses após aplicação) |
|-------|-------------------------------------|-----------------------------------------|------------------------------------|---------------------------------|
| 1     | 7,9                                 | 4                                       | 3,9                                | 8                               |
| 2     | 9,9                                 | 8                                       | 1,9                                | 20                              |
| 3     | 8,6                                 | 9                                       | 0,4                                | 8                               |
| 4     | 5,9                                 | 6                                       | 0,1                                | 20                              |
| 5     | 10,9                                | 6                                       | 4,9                                | 32                              |
| 6     | 5,3                                 | 3                                       | 2,3                                | 44                              |
| 7     | 8,9                                 | 9                                       | 0,1                                | 56                              |
| 8     | 14,0                                | 6                                       | 8,0                                | 80                              |
| 9     | 16,5                                | 6                                       | 10,5                               | 92                              |
| 10    | 8,3                                 | 8                                       | 0,3                                | 8                               |
| 11    | 2,9                                 | 4                                       | 1,1                                | 20                              |
| 12    | 4,0                                 | 4                                       | 0,0                                | 20                              |
| 13    | 5,9                                 | 6                                       | 0,1                                | 44                              |
| 14    | 16,0                                | 9                                       | 7,0                                | 80                              |
| 15    | 7,9                                 | 8                                       | 0,1                                | 8                               |
| 16    | 5,6                                 | 4                                       | 1,6                                | 8                               |

Ainda quanto a subdivisão G, a Regra (9) foi gerada a partir da base BD2, com os atributos época, CaCTCe<sub>AB</sub> e MgCTCe<sub>AB</sub> selecionados utilizando o método ACPS.

$$(9) \quad \text{gesso} = 0,0545 \times \text{época} + 0,137 \times \text{CaCTCe}_{AB} - 0,2296 \times \text{MgCTCe}_{AB} + 1,6416$$

A Regra (9) apresentou  $r = 0,66$  e  $\text{EMA} = 2,28$ , e os atributos que compuseram a regra foram os mesmos daqueles presentes na Regra (2), exceto os valores dos coeficientes dos atributos que foram diferentes e os valores de  $r$  e EMA que, no caso da Regra (9), foram piores.

Já, para a base BD3, ainda com a subdivisão G, as regras geradas apresentaram  $r$  abaixo de 0,70 e EMA acima de 1,85, e do ponto de vista agrônomo essas regras também não foram satisfatórias, ou seja, não deveriam ser utilizadas para a estimativa de gesso agrícola.

Em relação à subdivisão H, a base de dados BD1 gerou, em sua maioria, regras contendo condicionais. As regras no formato SE-ENTÃO, no caso da estimativa do gesso, não são recomendadas por ficarem restritas a uma faixa de valores para determinados atributos.

Na base BD2, ainda com subdivisão H, apesar de as regras geradas não apresentarem o formato SE-ENTÃO, apenas a Regra (10) se destacou

agronomicamente. A regra foi gerada a partir do conjunto de atributos compostos por época,  $CaCTCe_C$  e  $MgCTCe_C$  que foram selecionados pelo método ACPS.

$$(10) \text{ gesso} = 0,051 \times \text{época} + 0,2423 \times CaCTCe_C - 7,6888$$

A Regra (10), apesar de representar ótimas respostas para a estimativa de gesso, como pode ser observado no Quadro 6, apresentou valores de  $r$  (0,63) e EMA (2,50) considerados não satisfatórios computacionalmente.

Quadro 6 – Estimativa para aplicação de gesso utilizando a Regra (10)

| LINHA | ESTIMATIVA<br>(t ha <sup>-1</sup> ) | QTDE. APLICADA<br>(t ha <sup>-1</sup> ) | DIFERENÇA<br>(t ha <sup>-1</sup> ) | ÉPOCA<br>(meses após aplicação) |
|-------|-------------------------------------|-----------------------------------------|------------------------------------|---------------------------------|
| 1     | 4,1                                 | 4                                       | 0,1                                | 8                               |
| 2     | 7,7                                 | 8                                       | 0,3                                | 20                              |
| 3     | 9,6                                 | 9                                       | 0,6                                | 8                               |
| 4     | 4,4                                 | 6                                       | 1,6                                | 20                              |
| 5     | 9,0                                 | 6                                       | 3,0                                | 32                              |
| 6     | 4,2                                 | 3                                       | 1,2                                | 44                              |
| 7     | 8,0                                 | 9                                       | 1,0                                | 56                              |
| 8     | 8,4                                 | 6                                       | 2,4                                | 80                              |
| 9     | 11,8                                | 6                                       | 5,8                                | 92                              |
| 10    | 9,3                                 | 8                                       | 1,3                                | 8                               |
| 11    | 5,7                                 | 4                                       | 1,7                                | 20                              |
| 12    | 8,5                                 | 4                                       | 4,5                                | 20                              |
| 13    | 2,1                                 | 6                                       | 3,9                                | 44                              |
| 14    | 7,8                                 | 9                                       | 1,2                                | 80                              |
| 15    | 8,5                                 | 8                                       | 0,5                                | 8                               |
| 16    | 7,5                                 | 4                                       | 3,5                                | 8                               |

Já, na base BD3 com subdivisão H, houve regras com e sem condicionantes e, assim como nas bases BD1 e BD2, não foram recomendáveis para a estimativa de gesso agrícola. Além disso, os valores de  $r$  e EMA das regras geradas foram considerados baixos, sendo o  $r$  médio = 0,43 e EMA médio = 2,43, com desvio padrão de 0,02 e 0,04, respectivamente. Uma possível explicação para o baixo valor de  $r$  e a alto valor de EMA de tais regras é elas envolverem atributos somente das camadas mais profundas do solo.

Os valores de  $r$  e EMA médios das subdivisões G e H, para as bases BD1, BD2 e BD3, são mostrados na Tabela 12. Os valores de  $r$  e EMA são os valores médios obtidos a partir de cada execução do algoritmo *M5Rules*.

Tabela 12 – Comparativo das subdivisões G e H para as bases BD1, BD2 e BD3

| Superfície e Subsuperfície |              |      |              |      |
|----------------------------|--------------|------|--------------|------|
| BD1                        |              |      |              |      |
|                            | Subdivisão G |      | Subdivisão H |      |
|                            | $r$          | EMA  | $r$          | EMA  |
| <b>Média</b>               | 0,84         | 1,71 | 0,70         | 2,27 |
| <b>Desvio Padrão</b>       | 0,01         | 0,03 | 0,04         | 0,16 |
| BD2                        |              |      |              |      |
|                            | Subdivisão G |      | Subdivisão H |      |
|                            | $r$          | EMA  | R            | EMA  |
| <b>Média</b>               | 0,66         | 2,34 | 0,62         | 2,54 |
| <b>Desvio Padrão</b>       | 0,10         | 0,27 | 0,05         | 0,18 |
| BD3                        |              |      |              |      |
|                            | Subdivisão G |      | Subdivisão H |      |
|                            | $r$          | EMA  | $r$          | EMA  |
| <b>Média</b>               | 0,67         | 1,93 | 0,43         | 2,53 |
| <b>Desvio Padrão</b>       | 0,04         | 0,14 | 0,03         | 0,04 |

É possível perceber que nas três bases de dados BD1, BD2 e BD3, a subdivisão G apresentou os melhores resultados, tanto para o valor de  $r$  quanto para o de EMA. O resultado mais evidente foi constatado na base BD3, em que a média do valor de  $r$  foi 0,67 na subdivisão G e, na subdivisão H, a média do valor de  $r$  foi 0,43. Em todos os experimentos da segunda fase, pode-se constatar que as subdivisões que englobaram os atributos apenas da superfície apresentaram resultados computacionais melhores do que as subdivisões que englobaram apenas os atributos da subsuperfície.

Em relação à aplicação da técnica SMOTE na segunda fase dos experimentos, duas regras foram geradas semelhantes à Regra (9), exceto pelos valores dos coeficientes dos atributos. A primeira foi a Regra (11), cujos atributos que entraram na análise foram época,  $Ca_{AB}$ ,  $Mg_{AB}$ ,  $CaCTCe_{AB}$  e  $MgCTCe_{AB}$ , selecionados a partir do critério de seleção B2. Já, a segunda foi a Regra (12), que utilizou os atributos época,  $Mg_{AB}$ ,  $CaCTCe_{AB}$  e  $MgCTCe_{AB}$  selecionados a partir do método ACPS. Ambas as regras foram obtidas a partir da base BD1 e subdivisão C.

Apesar de os atributos envolvendo os teores terem sido analisados, o algoritmo *M5Rules* compôs as regras apenas com os atributos que continham o maior ganho de informação que, nesses casos, foram os atributos envolvendo saturação.

Os valores de  $r$  e EMA para a Regra (11) foram 0,62 e 2,40, respectivamente. Para a Regra (12), os valores de  $r$  (0,82) e EMA (1,71) foram mais satisfatórios. Isso ocorre porque o subconjunto de atributos utilizados na análise foi diferente.

$$(11) \quad \text{gesso} = 0,0677 \times \text{época} + 0,1575 \times \text{CaCTCe}_{AB} - 0,2815 \times \text{MgCTCe}_{AB} + 1,5963$$

$$(12) \quad \text{gesso} = 0,0727 \times \text{época} + 0,1494 \times \text{CaCTCe}_{AB} - 0,2925 \times \text{MgCTCe}_{AB} + 2,3616$$

### 4.1.3 Terceira fase

#### 4.1.3.1 Seleção de atributos com matriz de covariância S

Assim como na segunda fase, os atributos envolvendo somente o teor foram analisados apenas da base BD3 pelo fato de os dados serem de uma única área experimental – Área B. Então, para a subdivisão I, que envolve os teores dos cátions de todas as camadas (A, B, C e D), mais os atributos época e gesso, a análise dos resultados foi feita apenas na base BD3.

Nessas condições, a Regra (13) apresentou  $r = 0,71$  e  $\text{EMA} = 1,90$ . A regra foi gerada a partir dos atributos época,  $\text{Ca}_A$ ,  $\text{Ca}_B$ ,  $\text{Ca}_C$  e  $\text{Mg}_B$ , selecionados pelo critério B2.

$$(13) \quad \text{gesso} = 0,0485 \times \text{época} + 0,2579 \times \text{Ca}_A - 0,1783 \times \text{Mg}_B - 5,3371$$

Apesar de essa regra tratar o Ca e Mg em camadas diferentes, ambas são das camadas superficiais. Com isso, a Regra (13) apresentou ótimas estimativas, as quais podem ser visualizadas no Quadro 7. Em sete casos a diferença entre a dose estimada e quantidade real aplicada de gesso foi menor que  $0,8 \text{ t ha}^{-1}$ .

Quadro 7 – Estimativa para aplicação de gesso utilizando a Regra (13)

| LINHA | ESTIMATIVA<br>(t ha <sup>-1</sup> ) | QTDE. APLICADA<br>(t ha <sup>-1</sup> ) | DIFERENÇA<br>(t ha <sup>-1</sup> ) | ÉPOCA<br>(meses após aplicação) |
|-------|-------------------------------------|-----------------------------------------|------------------------------------|---------------------------------|
| 1     | 7,7                                 | 9                                       | 1,3                                | 8                               |
| 2     | 6,4                                 | 6                                       | 0,4                                | 20                              |
| 3     | 3,0                                 | 3                                       | 0,0                                | 20                              |
| 4     | 5,3                                 | 6                                       | 0,7                                | 32                              |
| 5     | 3,4                                 | 3                                       | 0,4                                | 44                              |
| 6     | 4,9                                 | 6                                       | 1,1                                | 80                              |
| 7     | 4,7                                 | 3                                       | 1,7                                | 80                              |
| 8     | 5,9                                 | 6                                       | 0,1                                | 92                              |
| 9     | 4,2                                 | 6                                       | 1,8                                | 8                               |
| 10    | 7,2                                 | 6                                       | 1,2                                | 20                              |
| 11    | 8,7                                 | 9                                       | 0,3                                | 20                              |
| 12    | 3,9                                 | 6                                       | 2,1                                | 44                              |
| 13    | 8,9                                 | 9                                       | 0,1                                | 92                              |
| 14    | 4,5                                 | 3                                       | 1,5                                | 56                              |
| 15    | 5,0                                 | 3                                       | 2,0                                | 20                              |
| 16    | 4,2                                 | 6                                       | 1,8                                | 8                               |

Nas bases BD1, BD2 e BD3, com subdivisão J, foram obtidas regras com ótimos valores de  $r$ , quando se consideraram apenas os atributos referentes à saturação de todas as camadas (A, B, C e D), mais os atributos época e gesso. Entretanto, do ponto de vista agrônomo, as regras geradas não foram indicadas para a estimativa de gesso agrícola.

O que dificultou a aplicabilidade das regras foi o fato de as subdivisões para a terceira fase envolverem os atributos de todas as camadas de solo. Apesar de existirem regras agronomicamente coerentes, como as Regras (14) e (15), essas seriam de difícil aplicação por envolverem o mesmo atributo em diferentes camadas. Ambas as regras, geradas a partir da base de dados BD3, foram coerentes no sentido dos atributos saturação por Mg e saturação por K terem sido inversamente proporcionais, enquanto que o atributo saturação por Ca foi diretamente proporcional, à dose de gesso.

$$(14) \text{ gesso} = 0,108 \times \text{época} - 0,2967 \times \text{MgCTCe}_A - 0,1204 \times \text{MgCTCe}_B - 0,339 \times \text{KCTCe}_C + 15,7802$$

$$(15) \text{ gesso} = 0,1186 \times \text{época} + 0,2663 \times \text{CaCTCe}_A + 0,0788 \times \text{CaCTCe}_C - 0,043 \times \text{MgCTCe}_D - 17,5739$$

Apesar da ausência de regras consideráveis no âmbito da estimativa do gesso nesses experimentos, é interessante considerar que, na base BD2, ao utilizar a seleção de atributos, tanto com o critério B2 quanto o B4, os atributos envolvendo a saturação por Ca estiveram presentes como um dos três atributos mais relevantes.

Ao utilizar o critério B2, os atributos mais relevantes foram  $m_B$ ,  $\text{CaCTCe}_A$  e  $\text{CaCTCe}_C$ , e com o critério B4, os atributos mais relevantes foram época,  $\text{CaCTCe}_C$  e  $\text{MgCTCe}_B$ . Isso evidencia a importância do atributo  $\text{CaCTCe}$  para a estimativa da dose de gesso, independentemente da camada de solo. Porém, por conterem dados de todas as camadas, as regras se tornaram de difícil aplicabilidade para a estimativa de gesso.

Com a subdivisão B, que envolveu tanto o teor quanto a saturação de cátions em todas as camadas (A, B, C e D), mais os atributos época e gesso, as bases BD1 e BD2 não apresentaram regras coerentes agronomicamente, apesar de terem sido geradas regras com valores de  $r$  e EMA satisfatórios. No entanto, na base de dados BD3, a Regra (16) foi considerada adequada para a estimativa do gesso. O interessante desta regra é o fato dela envolver apenas os atributos época, saturação por Ca na camada mais superficial (0–10 cm) e saturação por Ca na camada mais profunda (40–60 cm).

$$(16) \text{ gesso} = 0,0202 \times \text{época} + 0,2174 \times \text{CaCTCe}_A + 0,0503 \times \text{CaCTCe}_D - 12,1985$$

A Regra (16) apresentou  $r = 0,67$  e  $EMA = 1,98$ , e o conjunto de atributos utilizado para a geração do modelo para a estimativa de gesso foi composto pelos atributos época,  $K_D$ ,  $CaCTCe_A$  e  $CaCTCe_D$ , selecionados pelo critério de seleção B2. No entanto, o atributo  $K_D$  não compôs a regra por não apresentar ganho de informação, por isso, o próprio algoritmo *M5Rules* não utilizou o atributo na regra.

O Quadro 8 apresenta os resultados obtidos a partir da aplicação dessa regra em um subconjunto de teste. Apesar de terem sido obtidas diferenças entre a dose estimada e quantidade real aplicada de gesso de até  $5 \text{ t ha}^{-1}$ , houve nove casos em que essa diferença foi inferior a  $0,8 \text{ t ha}^{-1}$ .

Quadro 8 – Estimativa para aplicação de gesso utilizando a Regra (16)

| LINHA | ESTIMATIVA<br>( $\text{t ha}^{-1}$ ) | QTDE. APLICADA<br>( $\text{t ha}^{-1}$ ) | DIFERENÇA<br>( $\text{t ha}^{-1}$ ) | ÉPOCA<br>(meses após aplicação) |
|-------|--------------------------------------|------------------------------------------|-------------------------------------|---------------------------------|
| 1     | 9,0                                  | 4                                        | 5,0                                 | 8                               |
| 2     | 7,8                                  | 8                                        | 0,2                                 | 20                              |
| 3     | 8,8                                  | 9                                        | 0,2                                 | 8                               |
| 4     | 5,3                                  | 6                                        | 0,7                                 | 20                              |
| 5     | 8,6                                  | 6                                        | 2,6                                 | 32                              |
| 6     | 3,7                                  | 3                                        | 0,7                                 | 44                              |
| 7     | 7,2                                  | 9                                        | 1,8                                 | 56                              |
| 8     | 5,2                                  | 6                                        | 0,8                                 | 80                              |
| 9     | 7,5                                  | 6                                        | 1,5                                 | 92                              |
| 10    | 8,7                                  | 8                                        | 0,7                                 | 8                               |
| 11    | 5,4                                  | 4                                        | 1,4                                 | 20                              |
| 12    | 4,7                                  | 4                                        | 0,7                                 | 20                              |
| 13    | 4,7                                  | 6                                        | 1,3                                 | 44                              |
| 14    | 8,3                                  | 9                                        | 0,7                                 | 80                              |
| 15    | 8,3                                  | 8                                        | 0,3                                 | 8                               |
| 16    | 7,2                                  | 4                                        | 3,2                                 | 8                               |

#### 4.1.3.2 Seleção de atributos com matriz de covariância $\Omega''$

Os mesmos experimentos foram refeitos utilizando os critérios B2 e B4, a partir do cálculo da matriz de covariância, assumindo a independência entre os atributos da camada A em relação aos atributos da camada D.

As regras geradas a partir da subdivisão J, nas três bases (BD1, BD2 e BD3), não foram consideradas adequadas para a estimativa do gesso. Apesar disso, houve regras com coerência agronômica. A Regra (17) foi gerada a partir da base BD1, cujo valor de  $r$  (0,82) e  $EMA$  (1,77) foram satisfatórios. Essa regra foi obtida de um

subconjunto de atributos composto por  $m_B$ ,  $CaCTCe_C$ ,  $MgCTCe_A$ ,  $KCTCe_A$  e  $KCTCe_D$ , selecionados pelo critério B2.

$$(17) \quad gesso = 0,119 \times CaCTCe_C - 0,2687 \times MgCTCe_A - 0,1288 \times KCTCe_D + 5,8379$$

A regra foi considerada coerente agronomicamente, pois a saturação por Ca foi diretamente proporcional, e a saturação por Mg e por K, inversamente proporcional, à dose de gesso. Esse comportamento está coerente com os trabalhos de Quaggio *et al.* (1993), Caires *et al.* (1998) e Rampim, (2008). Porém, a regra apresenta os atributos em diferentes camadas de solo, dificultando a sua aplicação para a estimativa de gesso. Mesmo assim, cabe destacar que o atributo  $CaCTCe$  em diferentes camadas foi um dos três mais relevantes, tanto no critério B2 quanto no B4, e também quando se utilizou a matriz de covariância com independência. Isso reforça a importância do atributo  $CaCTCe$  para a estimativa da necessidade de gesso agrícola.

O experimento envolvendo a subdivisão I, na base BD3, agora utilizando a matriz de covariância, assumindo a independência entre os atributos das camadas A e D, gerou regras com coerência agrônômica, como a Regra (18).

$$(18) \quad gesso = 0,121 \times Ca_A + 0,3446 \times Ca_D - 0,121 \times Mg_C - 2,6925$$

A Regra (18) apresentou  $r = 0,71$  e  $EMA = 1,88$ , e foi gerada utilizando um subconjunto de atributos composto por  $Ca_A$ ,  $Ca_C$ ,  $Ca_D$ ,  $Mg_C$  e  $K_C$ , selecionados por meio do critério B2. A regra manteve a coerência de aumentar o teor de Ca e diminuir o teor de Mg, com o aumento das doses de gesso. Porém, o fato de se ter teores de camadas diferentes dificulta a aplicação da regra. De qualquer modo, o Quadro 9 apresenta as estimativas geradas com a aplicação da Regra (18). Apesar de a regra ter sido composta de teores de Ca e Mg em diferentes camadas, a Regra (18) apresentou ótimas estimativas, contendo nove casos em que a diferença entre a dose estimada e quantidade real aplicada foi igual ou menor do que  $1 \text{ t ha}^{-1}$  de gesso.

Quadro 9 – Estimativa para aplicação de gesso utilizando a Regra (18)

| LINHA | ESTIMATIVA<br>(t ha <sup>-1</sup> ) | QTDE. APLICADA<br>(t ha <sup>-1</sup> ) | DIFERENÇA<br>(t ha <sup>-1</sup> ) | ÉPOCA<br>(meses após aplicação) |
|-------|-------------------------------------|-----------------------------------------|------------------------------------|---------------------------------|
| 1     | 11,6                                | 9                                       | 2,6                                | 8                               |
| 2     | 5,5                                 | 9                                       | 3,5                                | 8                               |
| 3     | 7,6                                 | 6                                       | 1,6                                | 20                              |
| 4     | 3,6                                 | 3                                       | 0,6                                | 20                              |
| 5     | 7,0                                 | 6                                       | 1,0                                | 32                              |
| 6     | 3,9                                 | 3                                       | 0,9                                | 44                              |
| 7     | 8,4                                 | 9                                       | 0,6                                | 56                              |

|    |     |   |     |    |
|----|-----|---|-----|----|
| 8  | 4,2 | 6 | 1,8 | 80 |
| 9  | 5,7 | 6 | 0,3 | 92 |
| 10 | 5,1 | 6 | 0,9 | 8  |
| 11 | 5,7 | 6 | 0,3 | 20 |
| 12 | 8,2 | 9 | 0,8 | 20 |
| 13 | 4,3 | 6 | 1,7 | 44 |
| 14 | 7,6 | 9 | 1,4 | 92 |
| 15 | 3,4 | 3 | 0,4 | 56 |
| 16 | 5,1 | 6 | 0,9 | 8  |

Na subdivisão B, porém, com a utilização da matriz de covariância assumindo a independência entre os atributos das camadas A e D, não houve regras consideradas adequadas para a estimativa do gesso. Além disso, nesse caso, a maioria das regras geradas apresentou valor de  $r$  inferior a 0,70.

Após a análise das regras geradas foi possível observar dois pontos importantes nesses experimentos. O primeiro é quanto ao atributo K, tanto para o teor quanto para a saturação. A maioria das regras consideradas boas para a estimativa do gesso não envolveram este atributo, o qual foi descartado pela técnica de seleção de atributos ou pelo próprio algoritmo *M5Rules*. Porém, quando o atributo teor ou saturação por K foi incluído nas regras, os valores de  $r$  e EMA mantiveram-se semelhantes ou pioraram. As divergências quanto a esse nutriente podem ser encontradas em alguns trabalhos, como os de Silva *et al.* (1997) e Sousa e Ritchey (1986), nos quais constatou-se lixiviação de K trocável com a aplicação de gesso.

Já, nos trabalhos de Martins *et al.* (2002), Caires *et al.* (2004) e Serafim *et al.* (2012), não foram observadas alterações significativas no teor de K trocável com a adição de gesso. Em SPD, a lixiviação do K com o uso de gesso tem sido pequena, conforme constatado nos trabalhos de Caires *et al.* (1998, 2002). Desse modo, é possível afirmar que o K trocável (teor ou saturação) não é um atributo indicado para compor o modelo de estimativa de gesso, uma vez que a lixiviação desse nutriente pode ser dependente do tipo de solo.

O segundo ponto está relacionado aos critérios que têm sido empregados para a recomendação de gesso na agricultura. O gesso tem sido utilizado visando a melhoria do ambiente radicular das plantas e, por isso, todos os critérios atualmente empregados para a recomendação de gesso envolvem o teor de Ca, o teor de Al e a saturação por Al na camada subsuperficial do solo (RIBEIRO *et al.*, 1999). Não há nenhum critério de recomendação de gesso baseado em atributos da camada superficial.

Pode ser observado nos resultados desse trabalho que, em todos os experimentos da segunda fase, os resultados das análises que envolveram a subdivisão C, E e G (camada superficial do solo) foram superiores aos resultados obtidos a partir das análises das subdivisões D, F e H (camada subsuperficial do solo). Isso ocorreu tanto do ponto de vista numérico, para os valores de  $r$  e EMA, quanto do ponto de vista agrônomo, em que as regras geradas com as subdivisões C, E e G foram mais coerentes em relação às regras geradas pelas subdivisões D, F e G.

A melhoria ocasionada pelo gesso agrícola na camada superficial é tão ou mais importante que a melhoria causada na camada subsuperficial. No trabalho de Caires *et al.* (2004), em que houve aumento na produção de grãos do milho com o uso de gesso, os autores afirmaram que tal fato não foi ocasionado pelas alterações no crescimento do sistema radicular das plantas, e sim devido ao aumento da saturação por Ca nas camadas superficiais do solo. Esse resultado está de acordo com os obtidos no presente estudo, em que o atributo CaCTCe foi considerado como um dos mais relevantes, na maioria das seleções de atributos, para explicar a variação dos dados.

De acordo com Demattê (2011), ainda faltam informações para que a recomendação de gesso seja mais precisa. Esse é um assunto importante, pois o uso desordenado desse insumo pode causar desequilíbrio de nutrientes, principalmente de Mg e, em alguns casos, de K no solo. No entanto, como já comentado nas motivações desse trabalho, a recomendação de gesso com base na análise química de solo é de grande complexidade e ainda gera muitas dúvidas.

#### **4.2 Aplicação das melhores regras em um conjunto de testes**

Dentre todas as regras apresentadas na seção anterior, as Regras (4), (6), (10) e (16) destacaram-se por apresentar potencial na estimativa de gesso agrícola, tanto no âmbito agrônomo quanto computacional. O Quadro 10 apresenta as características dessas quatro regras.

Quadro 10 – Características das melhores regras geradas

| Regra | Fase | Base | Subdivisão | Critério de SA <sup>9</sup> | Atributos                                                         | Atributos CR <sup>10</sup>                       | <i>r</i> | EMA  |
|-------|------|------|------------|-----------------------------|-------------------------------------------------------------------|--------------------------------------------------|----------|------|
| (4)   | 2°   | BD2  | C          | ACPS                        | Época, Ca <sub>AB</sub> , CaCTCe <sub>AB</sub>                    | Época, CaCTCe <sub>AB</sub>                      | 0,63     | 2,42 |
| (6)   | 2°   | BD3  | E          | B2                          | Época, Ca <sub>AB</sub> , Mg <sub>AB</sub>                        | Época, Ca <sub>AB</sub> , Mg <sub>AB</sub>       | 0,71     | 1,86 |
| (10)  | 2°   | BD2  | H          | ACPS                        | Época, CaCTCe <sub>C</sub> , MgCTCe <sub>C</sub>                  | Época, CaCTCe <sub>C</sub>                       | 0,63     | 2,50 |
| (16)  | 3°   | BD3  | B          | B2                          | Época, K <sub>D</sub> , CaCTCe <sub>A</sub> , CaCTCe <sub>D</sub> | Época, CaCTCe <sub>A</sub> , CaCTCe <sub>D</sub> | 0,67     | 1,98 |

Os critérios de seleção que geraram os melhores conjuntos de atributos foi o B2 e o método ACPS, sendo os atributos época, Ca, Mg, CaCTCe e MgCTCe, independentemente da camada, os atributos mais importantes na estimativa do gesso. Mesmo assim, todas essas regras apresentaram valores de *r* e de EMA não satisfatórios. Essa característica comprova a importância do especialista na área porque tais regras, provavelmente, não teriam sido selecionadas por apresentarem tais valores de *r* e EMA. No entanto, no âmbito agrônomo, essas regras foram mais coerentes para a criação de modelos de estimativa da necessidade de gesso do que aquelas que apresentaram valores correlação (*r*) mais altos.

As quatro regras que se destacaram foram submetidas a uma nova aplicação, em um novo subconjunto de registros que foram selecionados levando em consideração aqueles que apresentaram produtividades relativas mais altas (acima de 80%). Os Quadros 11, 12, 13 e 14 apresentam as estimativas para as Regras (4), (6), (10) e (16), respectivamente.

Quadro 11 – Nova estimativa para aplicação de gesso utilizando a Regra (4)

| <b><i>gesso = 0,0664 × época + 0,2748 × CaCTCe<sub>AB</sub> - 12,714</i></b> |                                     |                                         |                                    |                                 |
|------------------------------------------------------------------------------|-------------------------------------|-----------------------------------------|------------------------------------|---------------------------------|
| LINHA                                                                        | ESTIMATIVA<br>(t ha <sup>-1</sup> ) | QTDE. APLICADA<br>(t ha <sup>-1</sup> ) | DIFERENÇA<br>(t ha <sup>-1</sup> ) | ÉPOCA<br>(meses após aplicação) |
| 1                                                                            | 3,5                                 | 4                                       | 0,5                                | 20                              |
| 2                                                                            | 5,1                                 | 8                                       | 2,9                                | 32                              |
| 3                                                                            | 5,3                                 | 6                                       | 0,7                                | 44                              |

<sup>9</sup> SA = Seleção de Atributos<sup>10</sup> CR = Atributos que compõe a regra

|    |      |    |     |    |
|----|------|----|-----|----|
| 4  | 5,5  | 9  | 3,5 | 68 |
| 5  | 9,5  | 9  | 0,5 | 68 |
| 6  | 7,9  | 9  | 1,1 | 80 |
| 7  | 13,4 | 9  | 4,4 | 92 |
| 8  | 9,1  | 8  | 1,1 | 8  |
| 9  | 10,5 | 12 | 1,5 | 20 |
| 10 | 10,8 | 12 | 1,2 | 20 |
| 11 | 6,6  | 9  | 2,4 | 44 |
| 12 | 4,4  | 6  | 1,6 | 8  |
| 13 | 4,3  | 4  | 0,3 | 8  |
| 14 | 2,2  | 4  | 1,8 | 20 |
| 15 | 5,0  | 3  | 2,0 | 20 |
| 16 | 8,3  | 12 | 3,7 | 8  |

Quadro 12 – Nova estimativa para aplicação de gesso utilizando a Regra (6)

| <b><math>gesso = 0,0396 \times epoca + 0,2891 \times Ca_{AB} - 0,2809 \times Mg_{AB} - 1,8504</math></b> |                                     |                                            |                                    |                                    |
|----------------------------------------------------------------------------------------------------------|-------------------------------------|--------------------------------------------|------------------------------------|------------------------------------|
| LINHA                                                                                                    | ESTIMATIVA<br>(t ha <sup>-1</sup> ) | QTDE.<br>APLICADA<br>(t ha <sup>-1</sup> ) | DIFERENÇA<br>(t ha <sup>-1</sup> ) | ÉPOCA<br>(meses após<br>aplicação) |
| 1                                                                                                        | 3,8                                 | 6                                          | 2,2                                | 8                                  |
| 2                                                                                                        | 4,0                                 | 6                                          | 2,0                                | 8                                  |
| 3                                                                                                        | 10,8                                | 9                                          | 1,8                                | 8                                  |
| 4                                                                                                        | 3,0                                 | 3                                          | 0,0                                | 20                                 |
| 5                                                                                                        | 5,8                                 | 6                                          | 0,2                                | 20                                 |
| 6                                                                                                        | 7,6                                 | 9                                          | 1,4                                | 20                                 |
| 7                                                                                                        | 4,8                                 | 6                                          | 1,2                                | 32                                 |
| 8                                                                                                        | 5,3                                 | 3                                          | 2,3                                | 44                                 |
| 9                                                                                                        | 5,2                                 | 3                                          | 2,2                                | 80                                 |
| 10                                                                                                       | 2,9                                 | 3                                          | 0,1                                | 80                                 |
| 11                                                                                                       | 5,6                                 | 9                                          | 3,4                                | 80                                 |
| 12                                                                                                       | 5,1                                 | 3                                          | 2,1                                | 80                                 |
| 13                                                                                                       | 1,3                                 | 3                                          | 1,7                                | 8                                  |
| 14                                                                                                       | 3,3                                 | 3                                          | 0,3                                | 20                                 |
| 15                                                                                                       | 2,7                                 | 3                                          | 0,3                                | 32                                 |
| 16                                                                                                       | 2,4                                 | 3                                          | 0,6                                | 44                                 |

Quadro 13 – Nova estimativa para aplicação de gesso utilizando a Regra (10)

| <b><math>gesso = 0,051 \times época + 0,2423 \times CaCTCe_c - 7,6888</math></b> |                                     |                                         |                                    |                                 |
|----------------------------------------------------------------------------------|-------------------------------------|-----------------------------------------|------------------------------------|---------------------------------|
| LINHA                                                                            | ESTIMATIVA<br>(t ha <sup>-1</sup> ) | QTDE. APLICADA<br>(t ha <sup>-1</sup> ) | DIFERENÇA<br>(t ha <sup>-1</sup> ) | ÉPOCA<br>(meses após aplicação) |
| 1                                                                                | 8,5                                 | 4                                       | 4,5                                | 20                              |
| 2                                                                                | 6,7                                 | 8                                       | 1,3                                | 32                              |
| 3                                                                                | 2,1                                 | 6                                       | 3,9                                | 44                              |
| 4                                                                                | 1,9                                 | 9                                       | 7,1                                | 68                              |

|    |      |    |     |    |
|----|------|----|-----|----|
| 5  | 8,4  | 9  | 0,6 | 68 |
| 6  | 5,0  | 9  | 4,0 | 80 |
| 7  | 13,1 | 9  | 4,1 | 92 |
| 8  | 9,0  | 8  | 1,0 | 8  |
| 9  | 9,4  | 12 | 2,6 | 20 |
| 10 | 9,1  | 12 | 2,9 | 20 |
| 11 | 6,0  | 9  | 3,0 | 44 |
| 12 | 7,6  | 6  | 1,6 | 8  |
| 13 | 4,3  | 4  | 0,3 | 8  |
| 14 | 4,4  | 4  | 0,4 | 20 |
| 15 | 5,4  | 3  | 2,4 | 20 |
| 16 | 8,1  | 12 | 3,9 | 8  |

Quadro 14 – Nova estimativa para aplicação de gesso utilizando a Regra (16)

| $\text{gesso} = 0,0202 \times \text{época} + 0,2174 \times \text{CaCTCe}_A + 0,0503 \times \text{CaCTCe}_D - 12,1985$ |                                     |                                            |                                    |                                    |
|-----------------------------------------------------------------------------------------------------------------------|-------------------------------------|--------------------------------------------|------------------------------------|------------------------------------|
| LINHA                                                                                                                 | ESTIMATIVA<br>(t ha <sup>-1</sup> ) | QTDE.<br>APLICADA<br>(t ha <sup>-1</sup> ) | DIFERENÇA<br>(t ha <sup>-1</sup> ) | ÉPOCA<br>(meses após<br>aplicação) |
| 1                                                                                                                     | 4,7                                 | 4                                          | 0,7                                | 20                                 |
| 2                                                                                                                     | 5,2                                 | 8                                          | 2,8                                | 32                                 |
| 3                                                                                                                     | 4,7                                 | 6                                          | 1,3                                | 44                                 |
| 4                                                                                                                     | 4,4                                 | 9                                          | 4,6                                | 68                                 |
| 5                                                                                                                     | 9,2                                 | 9                                          | 0,2                                | 68                                 |
| 6                                                                                                                     | 5,4                                 | 9                                          | 3,6                                | 80                                 |
| 7                                                                                                                     | 9,7                                 | 9                                          | 0,7                                | 92                                 |
| 8                                                                                                                     | 9,5                                 | 8                                          | 1,5                                | 8                                  |
| 9                                                                                                                     | 9,1                                 | 12                                         | 2,9                                | 20                                 |
| 10                                                                                                                    | 9,8                                 | 12                                         | 2,2                                | 20                                 |
| 11                                                                                                                    | 6,7                                 | 9                                          | 2,3                                | 44                                 |
| 12                                                                                                                    | 3,8                                 | 6                                          | 2,2                                | 8                                  |
| 13                                                                                                                    | 6,2                                 | 4                                          | 2,2                                | 8                                  |
| 14                                                                                                                    | 2,8                                 | 4                                          | 1,2                                | 20                                 |
| 15                                                                                                                    | 5,2                                 | 3                                          | 2,2                                | 20                                 |
| 16                                                                                                                    | 8,6                                 | 12                                         | 3,4                                | 8                                  |

Levando em consideração apenas as linhas que a diferença entre a quantidade aplicada e a quantidade estimada é menor ou igual a 2 t ha<sup>-1</sup>, tem-se que, para a Regra (4) apresentada no Quadro (11), a média da quantidade real aplicada é de 7 t ha<sup>-1</sup> e a média da diferença entre a quantidade aplicada e a quantidade estimada é de 1,10 t ha<sup>-1</sup>, representando então um erro de 15,71%. Para a Regra (6) apresentada no Quadro (12), a mesma análise resulta em uma média de 5,18 t ha<sup>-1</sup> da quantidade real aplicada, e uma

média de  $0,87 \text{ t ha}^{-1}$  da diferença entre a quantidade aplicada e quantidade estimada, o que representa um erro de 16,79%.

Para a estimativa da Regra (10) representada no Quadro (13), tem-se uma média de  $6,5 \text{ t ha}^{-1}$  da quantidade real aplicada e uma média de  $0,86 \text{ t ha}^{-1}$  da diferença entre a quantidade aplicada e a quantidade estimada, totalizando em 13% de erro.

Finalmente, para a Regra (16), apresentada no Quadro (14), esta análise tem uma média de  $6,66 \text{ t ha}^{-1}$  da quantidade real aplicada e uma média de  $0,93 \text{ t ha}^{-1}$  da diferença entre a quantidade aplicada e quantidade estimada, o que resulta em apenas 13,96% de erro.

### 4.3 Ampliação da base de dados com o método MTD

A aplicação do método MTD sobre as bases de dados agrícolas objetivou apenas a análise dos resultados computacionais quanto à melhoria ou não dos valores de  $r$  e EMA. Com isso, os resultados que serão apresentados nesta seção não envolvem a análise da coerência agronômica das regras obtidas a partir do uso do método MTD sobre as bases de dados.

Como a maioria das regras que foram consideradas viáveis agronomicamente faz parte da segunda fase dos experimentos, o método MTD foi aplicado sobre as bases de dados BD1, BD2 e BD3 nas subdivisões, C, D, E, F, G e H, utilizadas na segunda fase. O método MTD foi aplicado sobre o subconjunto de atributos que gerou as regras com maior  $r$  (melhor caso) e também sobre o subconjunto de atributos que gerou as regras com menor  $r$  (pior caso).

Os testes foram gerados utilizando os seguintes parâmetros: 10 registros para o conjunto de treinamento e 100 amostras virtuais geradas para cada subconjunto do atributo meta presentes no conjunto de treinamento (10-100), (20-100), (20-200), (20-400), (30-100), (30-300), (30-600), (40-100), (40-400), (40-800) e (40-1000).

Na maioria dos casos, em pelo menos um dos parâmetros utilizados para a aplicação do MTD, a resposta foi positiva, tanto no aumento do valor de  $r$  quanto na diminuição do valor de EMA. Exceto em três casos isso não ocorreu, sendo o primeiro na base BD1 com subdivisão C para o melhor caso, o segundo na base BD2 com subdivisão C para o melhor caso, e o terceiro foi apenas para o valor de  $r$  na base BD1 com subdivisão G, também no melhor caso.

Para representar os resultados obtidos, nas Tabelas 13, 14, 15, 16 e 17 estão apresentados os resultados da aplicação do método MTD nas bases BD1, BD2 e BD3 na

subdivisão G, que envolveu os atributos de saturação na superfície, mais os atributos época e gesso, tanto no melhor caso, quanto no pior caso. Em todas as tabelas constam os resultados obtidos com a utilização do método MTD e os valores de  $r$  e EMA são os valores médios de cada teste e não o valor de  $r$  e EMA de cada regra.

Tabela 13 – Aplicação do método MTD na base BD1 com subdivisão G

| BD1             |           |                                                 |        |      |        |                                               |        |      |       |
|-----------------|-----------|-------------------------------------------------|--------|------|--------|-----------------------------------------------|--------|------|-------|
| Atributos       |           | Melhor caso                                     |        |      |        | Pior caso                                     |        |      |       |
|                 |           | Época, MgCTCe <sub>AB</sub> , m <sub>AB</sub> , |        |      |        | CaCTCe <sub>AB</sub> , MgCTCe <sub>AB</sub> , |        |      |       |
| MTD             | Registros | $r$                                             | GANHO  | EMA  | GANHO  | $r$                                           | GANHO  | EMA  | GANHO |
|                 |           |                                                 | $r$    |      | EMA    |                                               | $r$    |      | EMA   |
| * <sup>11</sup> | 106       | 0,8                                             | (%)    | 1,67 | (%)    | 0,8                                           | (%)    | 1,73 | (%)   |
| <b>10-100</b>   | 5606      | 0,6                                             | -26,92 | 2,43 | 45,42  | 0,6                                           | -24,24 | 2,40 | 38,79 |
| <b>20-100</b>   | 3406      | 0,7                                             | -12,37 | 2,00 | 19,66  | 0,7                                           | -13,15 | 2,13 | 23,32 |
| <b>20-200</b>   | 7106      | 0,7                                             | -7,68  | 1,69 | 1,44   | 0,7                                           | -8,47  | 1,95 | 12,71 |
| <b>20-400</b>   | 13306     | 0,8                                             | -5,21  | 1,68 | 0,56   | 0,7                                           | -5,68  | 1,89 | 9,23  |
| <b>30-100</b>   | 2206      | 0,7                                             | -7,46  | 1,77 | 6,00   | 0,7                                           | -5,02  | 1,89 | 8,96  |
| <b>30-300</b>   | 6106      | 0,7                                             | -8,78  | 1,71 | 2,45   | 0,7                                           | -5,86  | 1,87 | 8,33  |
| <b>30-600</b>   | 12706     | 0,7                                             | -6,72  | 1,61 | -3,78  | 0,8                                           | -3,03  | 1,75 | 0,98  |
| <b>40-100</b>   | 1506      | 0,7                                             | -8,67  | 1,82 | 9,10   | 0,7                                           | -4,49  | 1,88 | 8,90  |
| <b>40-400</b>   | 5807      | 0,7                                             | -7,48  | 1,71 | 2,76   | 0,8                                           | -2,16  | 1,71 | -1,08 |
| <b>40-800</b>   | 11306     | 0,8                                             | -0,98  | 1,46 | -12,27 | 0,8                                           | 0,18   | 1,57 | -9,27 |
| <b>40-1000</b>  | 14106     | 0,7                                             | -6,11  | 1,65 | -0,93  | 0,8                                           | 1,17   | 1,57 | -9,02 |

A Tabela 13 apresenta os resultados obtidos a partir da base BD1. Pode-se observar que, no melhor caso, não houve aumento do valor de  $r$ , e que no pior caso, o aumento foi de apenas 1,17%. Uma possível explicação é que os valores de  $r$  sem a aplicação do método já foram considerados satisfatórios. Os ganhos nos valores de EMA para o melhor e o pior caso foram de 12,27% e 9,27%, respectivamente.

A Tabela 14 apresenta os resultados da aplicação do método MTD para a base BD2, também com a subdivisão G. Pode-se notar que, quando houve baixo valor de  $r$  e alto valor de EMA, os ganhos obtidos para tais índices foram maiores do que quando os valores de  $r$  e EMA já foram considerados satisfatórios. Tal fato foi mais evidente no pior caso, tendo ocorrido um ganho de 64,32% em relação ao valor de  $r$  e um ganho de 32% no valor de EMA. Já, no melhor caso, os ganhos não foram expressivos, mas o valor de EMA apresentou redução maior que 10%.

<sup>11</sup>\* = Sem a aplicação do método MTD

Tabela 14 – Aplicação do método MTD na base BD2 com subdivisão G

| <b>BD2</b> |           |                                                      |          |      |        |                                              |          |      |        |
|------------|-----------|------------------------------------------------------|----------|------|--------|----------------------------------------------|----------|------|--------|
| Atributos  |           | Melhor caso                                          |          |      |        | Pior caso                                    |          |      |        |
|            |           | Época, CaCTCe <sub>AB</sub> , MgCTCe <sub>AB</sub> , |          |      |        | Época, m <sub>AB</sub> , KCTCe <sub>AB</sub> |          |      |        |
| MTD        | Registros | <i>r</i>                                             | GANHO    | EMA  | GANHO  | <i>r</i>                                     | GANHO    | EMA  | GANHO  |
|            |           | <i>r</i>                                             | <i>r</i> | EMA  | EMA    | <i>r</i>                                     | <i>r</i> | EMA  | EMA    |
| *          | 144       | 0,7                                                  | (%)      | 2,22 | (%)    | 0,43                                         | (%)      | 2,96 | (%)    |
| 10-100     | 7744      | 0,5                                                  | -21,00   | 2,63 | 18,54  | 0,39                                         | -10,33   | 2,96 | -0,08  |
| 20-100     | 4844      | 0,6                                                  | -6,26    | 2,24 | 1,16   | 0,54                                         | 24,64    | 2,61 | -11,83 |
| 20-200     | 8944      | 0,6                                                  | -10,93   | 2,34 | 5,75   | 0,54                                         | 24,98    | 2,52 | -14,99 |
| 20-400     | 19344     | 0,6                                                  | -6,30    | 2,27 | 2,60   | 0,57                                         | 32,18    | 2,50 | -15,52 |
| 30-100     | 2944      | 0,7                                                  | -0,30    | 2,14 | -3,33  | 0,62                                         | 43,75    | 2,30 | -22,39 |
| 30-300     | 8544      | 0,7                                                  | 0,88     | 2,08 | -6,00  | 0,65                                         | 48,54    | 2,27 | -23,47 |
| 30-600     | 16944     | 0,7                                                  | 3,54     | 2,04 | -7,79  | 0,67                                         | 53,39    | 2,22 | -25,10 |
| 40-100     | 2244      | 0,7                                                  | 3,68     | 2,06 | -6,95  | 0,64                                         | 47,09    | 2,27 | -23,27 |
| 40-400     | 8544      | 0,7                                                  | 6,66     | 1,94 | -12,64 | 0,64                                         | 47,55    | 2,27 | -23,29 |
| 40-800     | 16944     | 0,7                                                  | 2,52     | 2,05 | -7,62  | 0,63                                         | 44,28    | 2,31 | -21,95 |
| 40-1000    | 21144     | 0,7                                                  | 5,27     | 1,97 | -11,05 | 0,71                                         | 64,32    | 2,01 | -32,04 |

Ainda na subdivisão G, a Tabela 15 apresenta os resultados obtidos com a aplicação do método MTD na base de dados BD3. No pior caso, o ganho sobre o valor de  $r$  foi superior a 10% e sobre o valor do EMA foi próximo a 10%. Já, no melhor caso, os ganhos sobre os valores de  $r$  e EMA não ultrapassaram 2%.

Tabela 15 – Aplicação do método MTD na base BD3 com subdivisão G

| <b>BD3</b> |           |                                                   |          |      |       |                                                   |          |      |       |
|------------|-----------|---------------------------------------------------|----------|------|-------|---------------------------------------------------|----------|------|-------|
| Atributos  |           | Melhor caso                                       |          |      |       | Pior caso                                         |          |      |       |
|            |           | Época, MgCTCe <sub>AB</sub> , KCTCe <sub>AB</sub> |          |      |       | Época, CaCTCe <sub>AB</sub> , KCTCe <sub>AB</sub> |          |      |       |
| MTD        | Registros | <i>r</i>                                          | GANHO    | EMA  | GANHO | <i>r</i>                                          | GANHO    | EMA  | GANHO |
|            |           | <i>r</i>                                          | <i>r</i> | EMA  | EMA   | <i>r</i>                                          | <i>r</i> | EMA  | EMA   |
| *          | 132       | 0,7                                               | (%)      | 1,86 | (%)   | 0,59                                              | (%)      | 2,20 | (%)   |
| 10-100     | 4732      | 0,6                                               | -9,85    | 2,12 | 14,24 | 0,56                                              | -3,69    | 2,27 | 3,03  |
| 20-100     | 2532      | 0,5                                               | -14,96   | 2,13 | 14,62 | 0,56                                              | -3,60    | 2,20 | -0,02 |
| 20-200     | 4932      | 0,6                                               | -3,63    | 1,94 | 4,70  | 0,62                                              | 5,19     | 2,14 | -2,72 |
| 20-400     | 9732      | 0,6                                               | -5,44    | 2,04 | 10,02 | 0,62                                              | 5,33     | 2,15 | -2,43 |
| 30-100     | 1732      | 0,6                                               | -7,98    | 2,05 | 10,42 | 0,57                                              | -2,78    | 2,25 | 2,35  |
| 30-300     | 4932      | 0,6                                               | -10,22   | 2,06 | 11,10 | 0,59                                              | 1,28     | 2,16 | -1,87 |
| 30-600     | 9732      | 0,6                                               | -9,36    | 2,06 | 11,07 | 0,58                                              | -0,79    | 2,21 | 0,30  |
| 40-100     | 1453      | 0,6                                               | -14,29   | 2,15 | 15,93 | 0,62                                              | 5,96     | 2,12 | -3,82 |
| 40-400     | 4932      | 0,6                                               | -5,14    | 1,97 | 6,12  | 0,62                                              | 5,18     | 2,08 | -5,45 |
| 40-800     | 9732      | 0,7                                               | 1,35     | 1,83 | -1,73 | 0,65                                              | 11,12    | 1,99 | -9,48 |
| 40-1000    | 12132     | 0,6                                               | -3,78    | 1,94 | 4,46  | 0,60                                              | 1,81     | 2,15 | -2,38 |

Nas Tabelas 13, 14 e 15 é possível observar que o maior ganho, tanto do valor de  $r$  quanto do valor de EMA, aconteceu quando houve 40 registros no conjunto de treinamento, gerando 400, 800 ou 1000 amostras virtuais. Com isso, pode-se concluir que quanto maior o número de registros utilizados para representar a base de dados,

melhor será o desempenho do método MTD. Os resultados na Tabela 16 mostram um caso da base de dados BD3 com subdivisão C, em que, no pior caso, o ganho do valor de  $r$  foi superior a 800%. Isso comprova que os ganhos obtidos ao se aplicar o método MTD são maiores quando os valores de  $r$  são mais baixos. Destaca-se que o valor de  $r$  antes da aplicação do método era 0,05 e, em todos os casos com a aplicação da técnica de ajuste de base de dados, houve aumento expressivo no valor de  $r$ , chegando a 827,98%. Porém, o valor de  $r$  ainda continuou baixo (0,47). Já, no melhor caso, o ganho do valor de  $r$  chegou, no máximo, a 3,54%.

Tabela 16 – Aplicação do método MTD na base BD3 com subdivisão C

| <b>BD3</b>     |           |                                         |        |      |       |                             |        |      |        |
|----------------|-----------|-----------------------------------------|--------|------|-------|-----------------------------|--------|------|--------|
| Atributos      |           | Melhor caso                             |        |      |       | Pior caso                   |        |      |        |
|                |           | Época, $Al_{AB}$ , $CA_{AB}$ , $K_{AB}$ |        |      |       | Época, $Al_{AB}$ , $K_{AB}$ |        |      |        |
| MTD            | Registros | $r$                                     | GANHO  | EMA  | GANHO | $r$                         | GANHO  | EMA  | GANHO  |
|                |           |                                         | $r$    |      | EMA   |                             | $r$    |      | EMA    |
| *              | 132       | 0,75                                    | (%)    | 1,69 | (%)   | 0,05                        | (%)    | 2,97 | (%)    |
| <b>10-100</b>  | 5232      | 0,59                                    | -22,27 | 2,26 | 33,79 | 0,31                        | 522,22 | 2,76 | -7,21  |
| <b>20-100</b>  | 2532      | 0,65                                    | -13,44 | 2,04 | 20,68 | 0,35                        | 596,43 | 2,69 | -9,38  |
| <b>20-200</b>  | 4932      | 0,66                                    | -12,67 | 2,00 | 18,75 | 0,37                        | 630,16 | 2,58 | -13,19 |
| <b>20-400</b>  | 9732      | 0,74                                    | -1,91  | 1,71 | 1,65  | 0,40                        | 684,13 | 2,59 | -12,99 |
| <b>30-100</b>  | 1732      | 0,71                                    | -5,32  | 1,87 | 10,94 | 0,30                        | 499,80 | 2,79 | -6,19  |
| <b>30-300</b>  | 4932      | 0,70                                    | -7,36  | 1,93 | 14,35 | 0,36                        | 607,94 | 2,64 | -11,24 |
| <b>30-600</b>  | 9732      | 0,73                                    | -3,06  | 1,75 | 3,91  | 0,46                        | 816,47 | 2,44 | -17,80 |
| <b>40-100</b>  | 4932      | 0,74                                    | -2,29  | 1,75 | 3,58  | 0,39                        | 673,61 | 2,58 | -13,17 |
| <b>40-400</b>  | 4932      | 0,73                                    | -3,77  | 1,73 | 2,70  | 0,41                        | 719,05 | 2,54 | -14,41 |
| <b>40-800</b>  | 9732      | 0,78                                    | 3,54   | 1,59 | -5,51 | 0,47                        | 827,98 | 2,43 | -18,44 |
| <b>40-1000</b> | 12132     | 0,74                                    | -1,45  | 1,71 | 1,66  | 0,46                        | 811,90 | 2,45 | -17,73 |

A Tabela 17 apresenta os resultados obtidos a partir da base BD1 com subdivisão F, nos quais o ganho em percentual não foi elevado, porém o valor de  $r$  passou de 0,71 a 0,91. Isso demonstra que o ganho, em percentual, deve sempre ser considerado de forma associada aos valores de correlação efetivamente obtidos.

Tabela 17 – Aplicação do método MTD na base BD3 com subdivisão F

| <b>BD1</b>    |           |                        |        |      |       |                                |       |      |        |
|---------------|-----------|------------------------|--------|------|-------|--------------------------------|-------|------|--------|
| Atributos     |           | Melhor caso            |        |      |       | Pior caso                      |       |      |        |
|               |           | Época, $Ca_C$ , $Mg_C$ |        |      |       | Época, $Ca_C$ , $Al_C$ , $K_C$ |       |      |        |
| MTD           | Registros | $r$                    | GANHO  | EMA  | GANHO | $r$                            | GANHO | EMA  | GANHO  |
|               |           |                        | $r$    |      | EMA   |                                | $r$   |      | EMA    |
| <b>Não</b>    | 106       | 0,77                   | (%)    | 2,01 | (%)   | 0,71                           | (%)   | 2,18 | (%)    |
| <b>10-100</b> | 5806      | 0,56                   | -27,25 | 2,54 | 26,64 | 0,71                           | 0,27  | 2,06 | -5,71  |
| <b>20-100</b> | 3506      | 0,69                   | -10,63 | 2,23 | 11,06 | 0,78                           | 10,00 | 1,83 | -16,00 |
| <b>20-200</b> | 6706      | 0,73                   | -5,09  | 2,07 | 3,33  | 0,84                           | 18,93 | 1,52 | -30,29 |
| <b>20-400</b> | 13306     | 0,65                   | -15,35 | 2,29 | 14,19 | 0,83                           | 18,04 | 1,50 | -31,04 |
| <b>30-100</b> | 2206      | 0,79                   | 3,04   | 1,81 | -9,92 | 0,81                           | 14,11 | 1,14 | -47,81 |

|                |       |      |        |      |        |      |       |      |        |
|----------------|-------|------|--------|------|--------|------|-------|------|--------|
| <b>30-300</b>  | 6406  | 0,63 | -17,99 | 2,26 | 12,33  | 0,83 | 16,70 | 1,54 | -29,20 |
| <b>30-600</b>  | 12706 | 0,77 | 0,61   | 1,69 | -15,59 | 0,87 | 22,88 | 1,27 | -41,80 |
| <b>40-100</b>  | 1506  | 0,80 | 4,42   | 1,63 | -19,00 | 0,87 | 23,59 | 1,27 | -41,74 |
| <b>40-400</b>  | 5706  | 0,75 | -2,57  | 1,91 | -4,74  | 0,86 | 21,83 | 1,33 | -38,86 |
| <b>40-800</b>  | 10506 | 0,82 | 6,57   | 1,56 | -22,50 | 0,91 | 28,66 | 0,92 | -57,59 |
| <b>40-1000</b> | 14106 | 0,79 | 3,27   | 1,68 | -16,14 | 0,89 | 25,30 | 1,15 | -47,26 |

Não se pode afirmar que ocorreu melhoria nos resultados de  $r$  e EMA em todos os casos em que houve aumento na quantidade de registros. Isso pode estar ligado ao fato de os melhores resultados terem sido gerados utilizando os parâmetros 40-400, 40-800 e 40-1000. Essa questão está ligada às características de qualidade *versus* o tamanho da base de dados. Por exemplo, na Tabela 17, tanto no melhor quanto no pior caso, ao utilizar o parâmetro 20-400 e 30-600, obteve-se 13306 e 12706 registros, respectivamente. No entanto, o melhor resultado, nesse caso, foi utilizando o parâmetro 40-800, o qual resultou em 10506 registros. Ou seja, o fato de os parâmetros 20-400 e 30-600 resultarem em uma quantidade de registros maior não significa necessariamente que o resultado foi melhor. Isso pôde ser observado na maioria dos testes realizados.

Outra característica observada é a quantidade de regras geradas, uma vez que, com a utilização do método MTD, o número de regras foi maior que nos experimentos sem aplicação do referido método. Por exemplo, utilizando a base BD1 com subdivisão G, ao definir o parâmetro como 20-400 (13306 registros), gerou-se mais de 40 regras. Isso pode estar associado ao fato do algoritmo utilizado, o *M5Rules*, que é aplicado de modo recursivo. Assim, todos os registros cobertos por uma regra vão sendo removidos da base até que todos os registros sejam cobertos por uma ou mais regras. Isso explica o motivo pelo qual, ao aumentar exponencialmente a base de dados, a quantidade de regras geradas também aumenta.

O trabalho de Duggal e Singh (2012) continha uma base de dados com 93 instancias e, apesar de ser considerada uma base de dados pequena, a resposta do algoritmo *M5Rules* foi a melhor dentre os algoritmos testados, podendo este ser indicado como o mais apropriado para a criação de um modelo no âmbito de estimativa de esforço de software.

Os trabalhos de Pina et al. (2010) e Pina e Pina Filho (2010), com 1030 e 2066 amostras, respectivamente, não apontaram o algoritmo *M5Rules* como sendo o melhor dentre todos os algoritmos testados, embora ele tenha apresentado bons resultados. Entretanto, em Pina et al. (2011), cuja base de dados analisada continha 188 amostras, o

algoritmo *M5Rules* mostrou-se o mais apropriado para o problema da estimativa da tenacidade Charpy de soldas metálicas.

Com isso, é possível inferir que o algoritmo *M5Rules* apresenta resultados melhores quando a base de dados é pequena. Esse é o caso do presente trabalho, pois as bases BD1, BD2 e BD3 continham 106, 144 e 132 registros, respectivamente, e foi possível gerar bons modelos no âmbito da estimativa da necessidade de gesso agrícola. A aplicação do algoritmo *M5Rules* sobre as bases de dados ajustadas com o método MTD resultou em ganhos tanto no valor de  $r$  quanto no valor de EMA, mesmo que isto tenha ocorrido em apenas alguns dos parâmetros testados.

## 5. CONCLUSÕES

O processo KDD possibilitou compreender o comportamento do gesso agrícola em relação aos atributos químicos de solos sob plantio direto e às quantidades recomendadas, por meio da seleção de atributos na etapa de pré-processamento dos dados, da aplicação do algoritmo *M5Rules* para a definição de modelos de estimativa de gesso na etapa de MD e do apoio de especialista na área para a validação dos resultados obtidos na etapa de MD pós-processamento.

Por meio de técnicas de seleção de atributos ficou evidenciada, para a definição dos modelos, a importância dos atributos: tempo após a aplicação de gesso (época), saturação por Ca na CTCe (CaCTCe) e saturação por Mg na CTCe (MgCTCe). Esses atributos, na maioria dos casos, estiveram entre os três mais relevantes das bases de dados analisadas, ou seja, foram aqueles que melhor explicaram o comportamento dos dados frente à aplicação de gesso agrícola. A utilização de técnicas de seleção de atributos foi fundamental na definição de modelos de estimativa, uma vez que elas tornaram os modelos menos específicos devido a redução da dimensionalidade da base de dados em função da seleção dos atributos mais relevantes contidos na mesma.

A utilização do algoritmo *M5Rules* possibilitou a geração de uma série de modelos para a estimativa da necessidade de gesso agrícola. A participação do especialista humano na área de domínio em todas as etapas desse trabalho foi de grande importância tendo sido possível assim, filtrar quatro modelos que poderiam ser aplicados à realidade agrícola, e a partir dos modelos gerados foi possível notar que o atributo saturação por Ca na CTCe (CaCTCe) se mostrou o mais indicado para a estimativa da dose de gesso agrícola, na maior parte dos casos estudados.

Os modelos gerados a partir dos atributos químicos que envolveram somente as camadas superficiais do solo mostraram-se mais interessantes, tanto do ponto de vista agrônomo, pela coerência dos modelos obtidos, quanto numérico, pelos valores de  $r$  e EMA mais satisfatórios. Esse resultado é importante porque os critérios atuais de recomendação de gesso são baseados exclusivamente em atributos das camadas do subsolo. Pelos resultados aqui obtidos e em confirmação com outros trabalhos já existentes, o efeito do gesso agrícola nas camadas superficiais é tão ou mais importante que o seu efeito no subsolo.

A utilização do algoritmo *M5Rules* se mostrou agronomicamente mais eficiente quando aplicado sobre uma base de dados considerada pequena. Isso porque ao utilizar o método MTD para a ampliação da base de dados, a quantidade de regras geradas pelo

algoritmo *M5Rules* foi maior, apresentando modelos com alta especificidade. Porém, independentemente da coerência das regras geradas com a utilização do método MTD, o objetivo proposto foi alcançado, pois foi possível observar melhoria nos valores de  $r$  e EMA com a utilização de uma base de dados contendo maior número de observações.

## 6. TRABALHOS FUTUROS

Para melhor compreensão dos resultados obtidos no presente estudo seria interessante a realização dos seguintes trabalhos futuros:

- Aplicar os quatro modelos selecionados nesse trabalho sobre uma nova base de dados com experimentação envolvendo o uso de gesso agrícola, de modo a verificar a aplicabilidade dos modelos;
- Utilizar técnicas de pós-processamento, de acordo com interesse das regras, para análise de sua coerência agrônômica a partir da utilização do método MTD;
- Utilizar outro algoritmo para geração de modelos de estimativa de gesso agrícola sobre as bases de dados que foram ampliadas com o método MTD.

## 7. REFERÊNCIAS BIBLIOGRÁFICAS

ADAMS, F.; LUND, Z. F. Effect os chemical activity of soil solution aluminum on cotton root penetration of acid subsoils. *Soil Science*, v. 101, p. 193-198, 1966.

ALBERG, D.; LAST, M.; NEUMAN, R.; SHARON, A. Induction of mean output prediction trees from continuous temporal meteorological data. *Journal of Applied Quantitative Methods*, v. 4, n. 4, p. 485-494, 2009.

ALCARDE, J.C. Contraditória, confusa e polêmica: é a situação do gesso na agricultura. *Informações Agronômicas*, Piracicaba, n. 41, p. 1-3, 1988.

ALMEIDA, L. F.; RODRIGUES, V. P. Mineração de dados agrícolas desbalanceados. 2011, 63f. Trabalho de Conclusão de Curso (Graduação em Bacharelado em Informática) – Universidade Estadual de Ponta Grossa, Ponta Grossa, 2011.

ANTUNES, J. F. G.; OLIVEIRA, S. R. M.; RODRIGUES, L. H. A. Mineração de dados para classificação das fases fenológicas da cultura da cana-de-açúcar utilizando dados do sensor modis e de precipitação. In: VIII Congresso Brasileiro de Agroinformática. Bento Gonçalves, 2011.

BAIR, E.; HASTIE, T.; PAUL, D.; TIBSHIRANI, R. Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, v. 101, p. 119-137, 2006.

BEHRENS, T.; SCHOLTEN, T. A comparions of data-mining techniques in predictive soil mapping. *Developments in soil science*, v. 31, p. 353-364, 2006.

BORKERT, C. M.; PAVANCT, M. A.; LANTMANN, A. F. Considerações sobre o uso de gesso na agricultura. *Informações Agronômicas*, Piracicaba, n. 40, p. 1-2, 1987.

BOSCHI, R. S.; OLIVEIRA, S. R. M.; ASSAD, E. D. Técnicas de mineração de dados para análise da precipitação pluvial decenal no Rio Grande do Sul. *Engenharia Agrícola*, v. 31, n. 6, 2011.

BUCENE, L. C.; RODRIGUES, L. H. A.; MEIRA, C. A. A. Mineração de dados climáticos para previsão de geada e deficiência hídrica para as culturas do café e da cana-de-açúcar para o Estado de São Paulo, Embrapa Informática Agropecuária – Documentos (INFOTECA-E), 2002. Disponível em: < <http://www.infoteca.cnptia.embrapa.br/handle/doc/8675> >. Acesso em: 25 nov. 2012.

BÜRKLE, P. Y. Um método de pós-processamento de regras de associação com base nas relações de dependência entre os atributos. 2006, 91f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal Fluminense, Niterói, 2006.

CAIRES, E. F.; CHUERI, W. A.; MADRUGA, E. F.; FIGUEIREDO A. Alterações de características químicas do solo e resposta da soja ao calcário e gesso aplicados na superfície em sistema de cultivo sem preparo do solo. *Revista Brasileira Ciência do Solo*, v. 22, n. 1, p. 27-34, 1998.

CAIRES, E. F.; FONSECA, A. F.; MENDES, J.; CHUERI, W. A.; MADRUGA E. F. Produção de milho, trigo e soja em função das alterações das características químicas do solo pela aplicação de calcário e gesso na superfície, em sistema de plantio direto. *Revista Brasileira Ciência do Solo*, v. 23, n. 2, p. 315-327, 1999.

CAIRES, E. F.; BANZATTO, D. A.; FONSECA, A. F. Calagem na superfície em sistema plantio direto. *Revista Brasileira Ciência do Solo*, v. 24, n. 1, p.161-169, 2000.

CAIRES, E. F.; FELDHAUS, I. C.; BLUM, J. Crescimento radicular e nutrição da cevada em função da calagem e aplicação de gesso. *Bragantia*, Campinas, v. 60, n. 3, p. 213-223, 2001.

CAIRES, E. F.; FELDHAUS, I. C.; BARTH, G.; GARBUIO, F. J. Lime and Gypsum application on the wheat crop. *Scientia Agricola*, Piracicaba, v. 59, n. 2, p. 357-365, 2002.

CAIRES, E. F.; BLUM, J.; BARTH, G.; GARBUIO, F. J.; KUSMAN, M. T. Alterações químicas do solo e resposta da soja ao calcário e gesso aplicados na implantação do sistema de plantio direto. *Revista Brasileira Ciência do Solo*, v. 27, n. 2, p. 275-286, 2003.

CAIRES, E. F.; KUSMAN, M. T.; BARTH, G.; GARBUIO, F. J.; PADILHA, J. M. Alterações químicas do solo e resposta do milho à calagem e aplicação de gesso. *Revista Brasileira Ciência do Solo*, v. 28, n. 1, p. 125-136, 2004.

CAIRES, E. F.; GARBUIO, F. J.; ALLEONI, L. R. F. A.; CAMBRI, M. A. Calagem superficial e cobertura de aveia preta antecedendo os cultivos de milho e soja em sistema plantio direto. *Revista Brasileira Ciência do Solo*, v. 30, n. 1, p. 87-98, 2006.

CAIRES, E. F.; MASCHIETTO, E. H. G.; GARBUIO, F. J.; CHURKA, S.; JORIS, H. A. W. Surface application of gypsum in low acidic Oxisol under no-till cropping system. *Scientia Agricola*, Piracicaba, v. 68, n. 2, p. 209-216, 2011.

CASTANHEIRA, L. G. Aplicação das técnicas de Mineração de dados em problemas de classificação de padrões. 2008, 95f. Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal Minas Gerais, Minas Gerais, 2008.

CRAMER, E. Estimation of the mean and the covariance matrix under a marginal independence assumption – an application of matrix differential calculus. *Linear Algebra and its Applications*, v. 288, p. 219-228, 1999.

CRUZ, J. C.; PEREIRA FILHO, I. A.; ALVARENGA, R. C.; GOTIJO NETO, M. M.; VIANA, J. H. M.; OLIVEIRA, M. F.; MATRANGOLO, W. J. R. Sistema de plantio direto. Embrapa Milho e Sorgo, ISSN 1679-012X Versão Eletrônica - 6ª edição, 2010. Disponível em: <[http://www.cnpms.embrapa.br/publicacoes/milho\\_6\\_ed/sisplantiodireto.htm](http://www.cnpms.embrapa.br/publicacoes/milho_6_ed/sisplantiodireto.htm)>. Acesso em: 20 nov. 2012.

DEMATTE, J. L. I. Ação do gesso e do calcário na relação cálcio/magnésio do solo e na produtividade da cana-de-açúcar. *Informações Agronômicas* - IPNI, n. 136, 2011.

DIAS, L. E. Uso do gesso como insumo agrícola. Comunicado Técnico. Embrapa, CNPDS, ISSN 0103-9407, n.7, p.1-6, 1992.

DIAS, M. M. Parâmetros na escolha de técnicas e ferramentas de mineração de dados. *Acta Scientiarum*, v. 24, n. 6, p. 1715-1725, 2002.

DOURADO, C. S.; ÁVILA, A. M. H.; OLIVEIRA, S. R. M. Regionalização da precipitação no estado da Bahia por meio de técnicas de mineração de dados. In: Congresso Brasileiro de Meteorologia, Anais... Gramado: UFRGS, 2012.

DUGGAL, H.; SINGH, P. Comparative Study of the Performance of M5-Rules Algorithm with Different Algorithm. *Journal of Software Engineering and Applications*, v. 5, n. 4, p. 270-276, 2012.

ERNANI, P. R.; RIBEIRO, M. S.; BAYER, C. Modificações químicas em solos ácidos ocasionadas pelo método de aplicação de corretivos da acidez e de gesso agrícola. *Scientia Agrícola*, v. 68, n. 4, p. 825-831, 2001.

FARIA, C. M. B.; COSTA, N. D.; FARIA A. F. Ação do calcário e gesso sobre características químicas do solo e na produtividade e qualidade do tomate e melão. *Horticultura Brasileira*, v. 21, n. 4, p. 615-619, 2003.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence*, v. 17, n. 3, p. 37-54, 1996.

FERREIRA, D. F. Estatística Multivariada. 1. ed. Lavras: UFLA, 2008.

GALVÃO, N. D.; MARIN, H. F. Técnicas de mineração de dados: uma revisão de literatura. *Acta Paulista de Enfermagem*, v. 22, n. 5, p. 686-690, 2009.

GUARIENTI, E. M.; CIACCO, C. F.; CUNHA, G. R.; DEL DUCA, L. J. A.; CAMARGO, C. M. O. Avaliação do efeito de variáveis meteorológicas na qualidade industrial e no rendimento de grãos de trigo pelo emprego de análise de componentes principais. *Ciência e Tecnologia de Alimentos*, v. 23, n. 3, p. 500-510, 2003.

GUIMARÃES, A. M.; VRIESMANN, L. M.; CANTERI, M. G.; CATANEO, A.; MOLIN, J. P. Desenvolvimento de um algoritmo genético para mineração de dados na agricultura de precisão. In: II Simpósio Internacional de Agricultura de Precisão, Anais... Viçosa, 2002.

GUIMARÃES, A. M. Aplicação de computação evolucionária na mineração de dados físico-químicos da água e do solo. 2005, 145f. Tese (Doutorado em Agronomia) - Faculdade de Ciências Agrônômicas da UNESP, Botucatu, 2005.

HOLMES, G.; HALL, M.; FRANK, E. Generating Rule Sets from Model Trees. In: Twelfth Australian Joint Conference on Artificial Intelligence, p. 1–12, 1999.

HUANG, C. Information diffusion techniques and small-sample problem. *International Journal of Information Technology e Decision Making*, v. 1, n. 2, p. 229-249, 2002.

HUANG, J. J.; MORAGA, C. A diffusio-neural-network for learning from small samples. *International Journal of Approximate Reasoning*, v. 35, n. 2, p. 137-161, 2004.

HUANG, J. J.; TZENG, G. H.; ONG, C. S. Two-stage genetic programming (2SGP) for the credit scoring model. *Applied Mathematics and Computation*, v. 174, n. 2, p. 1039-1053, 2006.

JOLLIFFE, I. T. Discarding Variables in a Principal Component Analysis. I: Artificial Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Blackwell Publishing for the Royal Statistical Society, v. 21, n. 2, p. 160-173, 1972.

JOLLIFFE, I. T. Principal Components Analysis. 2. ed. New York: Springer Verlag, 2002. 487 p. ISBN 0-387-95442-2.

KAMINSKI, J.; SANTOS, D. R.; GATIBONI, L. C.; BRUNETTO, G.; SILVA, L. S. Eficiência da calagem superficial e incorporada precedendo o sistema plantio direto em um argissolo sob pastagem natural. *Revista Brasileira Ciência do Solo*, v. 29, n. 4, p. 573-580, 2005.

KANTARDZIC, M. Data Mining Concepts, models, methods and algorithms. 1. ed. IEEE Press. 2003. 339 p. ISBN 0-471-22852-4.

KOTSIANTIS, S.; KOSTOULAS, A.; LYKOUDIS, S.; ARGIRIOU, A.; MENAGIAS, K. Using data mining techniques for estimating minimum, maximum and average daily temperature values. *International Journal of Engineering and Applied Sciences*, v.3, n. 5, p. 233-237, 2007.

LI, D. C.; WU, C. S.; TSAI, T. I.; LINA, Y. S. Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge. *Computer e Operations Research*, v. 34, n. 4, p. 966-982, 2007a.

LI, D. C.; HSU, H. C.; TSAI, T. ;LU, T. J.; HU, S. C. A new method to help diagnose cancers for small sample size. *Expert Systems with Applications*, v. 33, n. 2, p. 420-424, 2007b.

LI, D. C.; YEH, C. W. A non-parametric learning algorithm for small manufacturing data sets. *Journal Expert Systems with Applications: An Internation Journal*, v. 34, n. 1, p. 391-398, 2008.

LI, D. C.; LIU, C. W. A neural network weight determination model designed uniquely for small data set learning. *Journal Expert Systems with Applications: An Internation Journal*, v. 36, n. 6, p. 9853-9858, 2009.

LI, D. C.; LIU, C. W. Extending attribute information for small data set classification. *IEEE Transactions on Knowledge and Data Engineering*, 2010. No prelo.

LI, D. C.; LIU, C. W.; HU, S. C. A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medicine*,. v. 40, n. 5,p. 509-518, 2010.

LIAO, T. W. Diagnosis of bladder cancers with small sample size via feature selection. *Expert Systems with Applications*, v. 38, n. 4, p. 4649-4654, 2011.

LOPES, A. S.; SILVA, M. C.; GUILHERME, L. R. G. Acidez do solo e calagem. *Boletim Técnico - Associação Nacional para Difusão de Adub*, São Paulo, n. 1, p. 5-8, 1991.

LOPES, A. S.; WIETHÖLTER, S.; GUILHERME, L. R. G.; SILVA, C. A. Sistema de plantio direto: Bases para o manejo da fertilidade do solo. *Livro Técnico - Associação Nacional para Difusão de Adubos*, São Paulo, p. 6-26, 2004.

MACHADO, E. L. Um estudo de limpeza em base de dados desbalanceada e com sobreposição de classes. 2007, 63f. Dissertação (Mestre em Informática) - Universidade de Brasília, Brasília,2007.

MARTINS, O. C.; NOVAIS, R. F.; ALVAREZ, V. H.; RIBEIRO, A. C.; BARROS, N. F. Respostas à aplicação de diferentes misturas de calcário e gesso em solos. I. alterações químicas no perfil do solo. *Revista Ceres*, v. 49, p. 123-135, 2002.

MARQUES, R. R. Aplicação superficial de calcário e gesso em manejo conservacionista de solo para cultivo de amendoim e a aveia branca. 2008, 158f. Tese (Doutorado em Agronomia) – Faculdade de Ciências Agronômicas da UNESP, Botucatu, 2008.

MASCARENHAS, H. A. A.; MIYASAKA, S.; IGUE, T.; FREIRE, E. S. Respostas da soja à calagem e a adubações minerais com fósforo e potássio em solo Latossolo roxo. *Bragantia*, v. 28, n. único, p. 17-21, 1969.

MASCARENHAS, H. A. A.; GALLO, J. R.; RAIJ, B. van, IGUE, T. Efeitos da calagem nas características químicas do solo e na nutrição de soja em Latossolo roxo distrófico de cerrado. *Bragantia*, v. 35, n. 24, p. 276-278, 1976.

MASCARENHAS, H. A. A.; BRAGA, N. R.; MIRANDA, M. A. C.; TISSELLI-FILHO, O.; MIYASAKA, S. Calagem e adubação da soja. 3.ed. In: A soja no Brasil Central. Campinas, Fundação Cargill, 1986. CD-ROM

MATHIAS, I. V. Aplicação de redes neurais artificiais na análise do molhamento foliar por orvalho. 2006, 144f. Tese (Doutorado em Agronomia) – Faculdade de Ciências Agrônômicas da UNESP, Botucatu, 2006.

MEIRA, C. A. A.; RODRIGUES, L. H. A. Mineração de padrões sequenciais climáticos para uso em sistema de previsão de doenças e danos na cultura do café no estado de São Paulo. In: V Congresso Brasileiro de Agroinformática, Porto Seguro, 2003.

MEIRA, C. A. A.; RODRIGUES, L. H. A.; MORAES, S. A. Análise da epidemia da ferrugem do cafeeiro com Árvore de Decisão. *Tropical Plant Pathology*, v. 33, n. 2, p. 114-124, 2008.

MOLIN, J. P.; COUTO, H. T. Z.; GIMENEZ, L. M.; PAULETTI, V.; MOLIN, R.; VIEIRA, S. R. Regression and correlation analysis of grid soil data versus cell spatial data. In: European Conference on Precision Agriculture, p. 449-453, 2001.

MUCHERINO, A.; PAPAJOJGI, P.; PARDALOS, P. M. A survey of data mining techniques applied to agriculture. *Operational Research An International Journal*, v. 9, n. 2, p. 121-140, 2009.

MUZILLI, O. Princípios e perspectivas de expansão. In: INSTITUTO AGRONÔMICO DO PARANÁ. Plantio direto no estado do Paraná. Londrina, p. 11-16, 1981. (Circular, 23)

NAVEGA, S. Princípios Essenciais do data mining. In: Infoimagem. Anais... Cenadem, 2002.

NEIS, L. Gesso agrícola em sistemas de manejo do solo e produtividade de soja na região do sudoeste de Goiás. 2009, 49f. Dissertação (Mestrado em Agronomia) - Universidade Federal de Goiás, Jataí, 2009.

NEIS, L.; PAULINO, H. B.; SOUZA, E. D.; REIS, E. F.; PINTO, F. A. Gesso agrícola e rendimento de grãos de soja na região do sudoeste de Goiás. *Revista Brasileira Ciência do Solo*, v. 34, n. 2, p. 409-416, 2010.

NG, R. T.; HAN, J. Efficient and effective clustering methods for spatial data mining. In: 20th International Conference on Very Large Data Bases, Chile, p. 144-155, 1994.

NOGUEIRA, M. A.; MELO, W. J. Enxofre disponível para a soja e atividade de arilsulfatase em solo tratado com gesso agrícola. *Revista Brasileira Ciência do Solo*, v. 27, n. 4, p. 655-663, 2003.

OLIVEIRA, J. A.; DUTRA, L. V.; RENNO, C. D. Aplicação de métodos de extração e seleção de atributos para classificação de regiões. In: XII Simpósio Brasileiro de Sensoriamento Remoto. Anais... Goiânia, p. 4201-4208, 2005.

OLIVEIRA, E. L.; PAVAN, M. A. Control of soil acidity in no-tillage system for soybean production. *Soil and Tillage Research*, v. 38, n. 1-2, p. 47-57, 1996.

OLIVEIRA Jr., J. I. Seleção de variáveis na mineração de dados agrícolas: Uma abordagem baseada em análise de componentes principais. 2012, 87f. Dissertação (Mestrado em Computação Aplicada) – Universidade Estadual de Ponta Grossa, Ponta Grossa, 2012.

PAPPA, G. L. Seleção de atributos utilizando algoritmos genéticos multiobjetivos. 2002, 75f. Dissertação (Mestrado em Informática Aplicada) – Universidade Católica do Paraná, Curitiba, 2002.

PAVAN, M. A.; BLOCH, M. F.; ZEMPULSKI, H. C.; MIYAZAWA, M.; ZOCOLER, D. C. 1992. Manual de análise química do solo e controle de qualidade. Instituto Agrônomo do Paraná (IAPAR), Londrina, 38 p.

PEEL, M. C.; FINLAYSON, B. L.; McMAHON, T. A. Update world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences*, v. 11, n. 5, p. 1633–1644, 2007.

PIMENTEL, D. P.; OMAR, N. Descobrendo conhecimentos em dados de avaliação da aprendizagem com técnicas de mineração de dados. In: XII Workshop de Informática na Escola. Anais... Campo Grande, p. 147-155, 2006.

PINA, A. C.; PINA FILHO, A. C. Aplicação do aprendizado de máquina na previsão dos resultados de ensaios de fluência em aços. In: Nono Simpósio de Mecânica Computacional. Universidade Federal de São João Del-Rei, 2010.

PINA, A. C.; GOMES, L.; RANGEL, C.; PINA FILHO, A. C. Aplicação de aprendizado de máquina na estimação da força compressiva do concreto. In: Nono Simpósio de Mecânica Computacional. Universidade Federal de São João Del-Rei, 2010.

PINA, A. C.; BARROS, L. R. S.; SOUZA, L. R.; PINA FILHO, A. C. Aplicando algoritmos de aprendizagem de máquinas na caracterização da qualidade da soldagem. In: 6º Congresso Brasileiro de engenharia de fabricação, Caxias do Sul, 2011.

PIRAMUTHU, S. Evaluating Feature Selection Methods for Learning in Data Mining Applications. *European Journal of Operational Research*, v. 156, n. 2, p. 483-494, 2004.

QUAGGIO, J. A.; RAIJ, B. van; GALLO, P. B.; MASCARENHAS, H. A. A. Respostas da soja à aplicação de calcário e gesso e lixiviação de íons no perfil do solo. *Pesquisa Agropecuária Brasileira*, v. 28, n. 3, p. 375-383, 1993.

RAIJ, B. van; CAMARGO, A. P.; CANTARELLA, H.; SILVA, N. M. Alumínio trocável e saturação em bases como critérios para recomendação de calagem. *Bragantia*, v. 42, n. 1, p. 149-156, 1983.

RAIJ, B. van. Fertilidade do solo e adubação. 1. ed. Editora Agronômica Ceres, 1991.

RAIJ, B. van; FURLANI, P. R.; QUAGGIO, J. A.; PATTINELLI JÚNIOR, A. P. Gesso na produção de cultivares de milho com tolerância diferencial a alumínio em três níveis de calagem. *Revista Brasileira Ciência do Solo*, v. 22, n. 1, p. 101-108, 1998.

RAIJ, B. van. Gesso na agricultura. *Informações Agronômicas*, n. 122, p. 26-27, 2008.

RAIJ, B. van. Melhorando o ambiente radicular em subsuperfície. *Informações Agrônomicas*, n. 135, p. 9-18, 2011.

RAMOS, L. A.; NOLLA, A.; KORNDÖRFER, G. H.; PEREIRA, H. S.; CAMARGO, M. S. Reatividade de corretivos da acidez e condicionadores de solo em colunas de lixiviação. *Revista Brasileira Ciência do Solo*, v. 30, n. 5, p. 849-857, 2006.

RAMPIM, L. Atributos químicos de um Latossolo Vermelho eutroférico submetido a gessagem e cultivado com trigo e soja em semeadura direta. 2008, 81f. Dissertação (Mestrado Agronomia) - Universidade Estadual do Oeste do Paraná, Marechal Cândido Rondon, 2008.

REEVE, N.G.; SUMNER, M.E. Amelioration of subsoil acidity in Natal Oxisols by leaching of surface applied amendments. *Agrochemophysica*, v. 4, p. 1-6, 1972.

RIBEIRO, A. C.; GUIMARÃES, P. T. G.; ALVAREZ, V. H. Recomendações para o uso de corretivos e fertilizantes em Minas Gerais. Comissão de Fertilidade do Solo do Estado de Minas Gerais - CFSEMG, 1999.

RIOS, M. A.; PEARSON, R. W. The effect of some chemical environmental factors on cotton root behavior. *Soil Science Society of America Proceedings*, v. 28, p. 232-235, 1964.

RITCHEY, K. D.; SOUSA, D. M. G.; LOBATO, E.; CORREA, O. Calcium leaching to increase rooting depth in a Brazilian Savannah Oxisol. *Agronomy Journal*, v. 72, p. 40-44, 1980.

RODRIGUES, C. F. Análise comparativa entre os métodos decomposição em valores singulares e análise de componentes principais envolvendo matrizes esparsas de grande porte. 2011, 171f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Minas Gerais, 2011.

RUSSELL, S. J.; NORVIG, P. Artificial Intelligence: A Modern Approach. 1. ed. Prentice Hall, 2004.

SANTOS, J. C.; OLIVEIRA, J. R. F.; DUTRA, L. V.; SANT'ANNA, S. J. S.; RENNÓ, C. D. Seleção de atributos usando algoritmos genéticos para classificação de regiões. In: XIII Simpósio Brasileiro de Sensoriamento Remoto. Anais... Florianópolis, p. 6143-6150, 2007.

SARABJOT, S. A.; BELL, D. A.; HUGHES, J. G. The role of domain knowledge in data mining. In: Fourth International Conference on Information and Knowledge Management, p. 37-43, 1995.

SERAFIM, M. E.; LIMA, J. M.; LIMA, V. M. P.; ZEVIANI, W. M.; PESSONI, P. T. Alterações físico-químicas e movimentação de íons em Latossolo gibbístico sob doses de gesso. *Bragantia*, v.71, n.1, p.75-81, 2012.

SILVA, N. M.; RAIJ, B. van; CARVALHO, L. H.; BATAGLIA, O. C.; KONDO, J. I. Efeitos do calcário e do gesso nas características químicas do solo e na cultura do algodão. *Bragantia*, v. 56, n. 2, p. 389-401, 1997.

SILVA, S. H. M. G.; CARVALHO, A. C. P. L. F.; ROMERO, R. A. F.; CRUVINEL, P. E.; NATALE, W. Redes neurais e agricultura de precisão para recomendação de adubação da cultura da goiabeira. *Revista Brasileira de Agrocomputação*, v. 2, n. 1, p. 37-42, 2004.

SILVA, I. A. F. Descoberta de conhecimento em base de dados de monitoramento ambiental para avaliação da qualidade da água. 2007, 119f. Dissertação (Mestrado em Física e Meio Ambiente) - Universidade Federal de Mato Grosso, 2007.

SILVA, L. L.; COSTA, R. F.; CAMPOS, R. J. H. B. C.; DANTAS, R. T. Influência das precipitações na produtividade agrícola no Estado da Paraíba. *Revista Brasileira de Engenharia Agrícola Ambiental*, v. 13, n. 4, p. 454-461, 2009.

SILVA, R.; GHAHRAMANI, Z. Factorial mixture of gaussians and the marginal independence model. In: 12<sup>th</sup> International Conference on Artificial Intelligence and Statistics, Florida, p. 520-527, 2009.

SILVA, S. A.; LIMA, J. S. S.; XAVIER, A. C.; TEIXEIRA, M. M. Variabilidade espacial de atributos químicos de um Latossolo vermelho-amarelo húmico cultivado com café. *Revista Brasileira de Ciência do Solo*, v. 34, n. 1, p. 15-22, 2010.

SODRÉ, G. S.; MARCHINI, L. C.; MORETI, A. C. C. C.; OTSUK, I. P.; CARVALHO, C. A. L. Caracterização físico-química de amostras de méis de *Apis mellifera* L. (Hymenoptera:Apidae) do Estado do Ceará. *Ciência Rural*, v. 37, n. 4, p. 1139-1144, 2007.

SORATTO, R. P.; CRUSCIOL, C. A. C. Atributos químicos do solo decorrentes da aplicação em superfície de calcário e gesso em sistema plantio direto recém-implantado. *Revista Brasileira Ciência do Solo*, v. 32, n. 2, p. 675-688, 2008a.

SORATTO, R. P.; CRUSCIOL, C. A. C. Métodos de determinação de cálcio e magnésio trocáveis e estimativa do calcário residual em um latossolo submetido à aplicação de calcário e gesso em superfície. *Revista Brasileira de Ciência do Solo*, v. 32, n. 2, p. 663-673, 2008b.

SOUSA, D. M. G.; RITCHEY, K. D. Uso do gesso no solo de cerrado. In: Seminário sobre o uso de fosfogesso na agricultura. Anais... Brasília, p.119-144, 1986.

SOUSA, D. M. G.; LOBATO, E.; REIN, T. A. Uso do gesso agrícola nos solos do cerrado. Documentos – Embrapa Cerrados, 2005.

SOUZA, Z. M.; CERRI, D. G. P.; COLET, M. J.; RODRIGUES, L. H. A.; MAGALHÃES, P. S. G.; MANDONI, R. J. A. Análise dos atributos do solo e da produtividade da cultura da cana-de-açúcar com uso da geoestatística e Árvore de Decisão. *Ciência Rural*, v. 40, n. 4, p. 840-847, 2010.

TAN, P. N.; STEINBACH, M.; KUMAR, V. Introdução ao DATAMINING Mineração de Dados. 1. ed.Editora Ciência Moderna, Rio de Janeiro, 2009.

TSAL, T. I.; LI, D. C. Utilize bootstrap in small data set learning for pilot run modeling of manufacturing systems. *Journal Expert Systems with Applications: An International Journal*, v. 35, n. 3, p. 1293-1300, 2008.

UNO, Y.; PRASHER, S. O.; LACROIX, R.; KARIMI, Y.; VIAU, A.; PATEL, R. M. Artificial neural networks to predict corn yield from compact airborne spectrographic imager data. *Computers and Electronics in Agriculture*, v. 47, n. 2, p. 149-161, 2005.

VALLADARES, G. S.; GOMES, E. G.; MELLO, J. B. C. C. S.; PEREIRA, M. G.; ANJOS, L. G. C.; EBELING, A. G.; BENITES, V. M. Análise de componentes principais e métodos multicritérios ordinais no estudo de Organossolos e solos afins. *Revista Brasileira de Ciência do Solo*, v. 32, n. 1, p. 285-296, 2008.

VELOSO, C. A. C.; BORGES, A. L.; MUNIZ, A. S.; VEIGAS, I. A. de J. M. Efeitos de diferentes materiais no pH do solo. *Scientia Agricola*, v. 49, n. 1, p. 123-128, 1992.

VRIESMANN, L. M.; GUIMARÃES, A. M.; CANTERI, M. G.; MOLIN, J. P.; CATANEO, A.; KOVALECHYN, D. Análise de resultados obtidos por técnicas de inteligência artificial da mineração de dados de produtividade do solo. *Revista Brasileira de Agrocomputação*, v. 2, n. 1, p. 11-18, 2004.

WEBER, A. R. H.; Oliveira Jr. J.I.; PROENÇA, C. A.; GUIMARÃES, A. M.; ROCHA, J. C. F.; POZO, A. T. R. Aplicação de Redes neurais artificiais para avaliação da influência da compactação do solo na produtividade de milho em dados agrícolas georreferenciados. In: VIII Congresso Brasileiro de Agroinformática, Bento Gonçalves 2011.

WILLIAMS, G.; BAXTER, R.; HE, H. e HAWKIS, S. A comparative study of RNN for outlier detection in data mining. In: IEEE International Conference on Data Mining, p. 709-712, 2002.

WIESEL, A.; HERO, A. O. III. Decomposable principal component analysis. *IEEE Transactions on signal processing*, v. 57, n. 11, p. 4369-4377, 2009.

WITTEN, I. H.; FRANK, E. Data Mining: Practical machine learning tools and techniques. 2. ed., Morgan Kaufmann, San Francisco, 2005.

YANG, T.; KECMAN, V. Adaptative local hyperplane algorithm for learning small medical data sets. *Expert Systems: The Journal of Knowledge Engineering*, v. 26, n. 4, p. 355-359, 2009.

ZHU, R. M. H.; ZHANG, L.; CHEN, A. A new method to assist small data set neural network learning. In: Sixth International Conference on Intelligent Systems Design and Applications, p.17-22, 2006.

ZIGLIO, C. M.; MIYAZAWA, M.; PAVAN, M. A. Formas orgânicas e inorgânicas de mobilização do cálcio no solo. *Brazilian Archives of Biology and Technology*, v. 42, n. 2, p. 257-262, 1999.