UNIVERSIDADE ESTADUAL DE PONTA GROSSA
MESTRADO EM COMPUTAÇÃO APLICADA

VALTER LUÍS ESTEVAM JUNIOR

INTERESSABILIDADE DE MODELOS DE REGRESSÃO EM MINERAÇÃO DE DADOS AGRÍCOLAS

PONTA GROSSA
2015

VALTER LUÍS ESTEVAM JUNIOR

INTERESSABILIDADE DE MODELOS DE REGRESSÃO EM MINERAÇÃO DE DADOS AGRÍCOLAS

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.
Orientadora: Profª Dra. Alaine Margarete Guimarães
Co-orientador: Profº Dr. Eduardo Fávero Caires

PONTA GROSSA
2015

Ficha Catalográfica
Elaborada pelo Setor de Tratamento da Informação BICEN/UEPG

Estevam Junior, Valter Luís
E79 Interessabilidade de modelos de regressão em mineração de dados agrícolas/ Valter Luís Estevam Junior. Ponta Grossa, 2015.
82f.

Dissertação (Mestrado em Computação Aplicada - Área de Concentração: Computação para Tecnologias em Agricultura), Universidade Estadual de Ponta Grossa.
Orientadora: Profª Drª Alaine Margarete Guimarães.
Coorientador: Prof. Dr. Eduardo Fávero Caires.

1.Gesso agrícola. 2.Redes bayesianas. 3.Sensitividade. 4.Impacto. I.Guimarães, Alaine Margarete. II. Caires, Eduardo Fávero. III. Universidade Estadual de Ponta Grossa. Mestrado em Computação Aplicada. IV. T.

CDD: 631.3

TERMO DE APROVAÇÃO

Valter Luis Estevam Junior

"INTERESSABILIDADE DE MODELOS DE REGRESSÃO EM MINERAÇÃO DE DADOS AGRÍCOLAS"

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:

Orientadora: *Drª Alaine Margarete Guimarães*
UEPG

Drª Aurora Trinidad Ramirez Pozo
UFPR

Dr. Fernando José Garbuio
IFC

Ponta Grossa, 26 de fevereiro de 2015.

Dedico este trabalho a Dulce, Valter, Fábio e Emanuele,

meus amores.

AGRADECIMENTOS

A Deus por ter me dado saúde e perseverança para seguir adiante com este projeto mesmo diante de tantas adversidades.

À minha mãe, Dulce, pelos seus ensinamentos e seu infinito amor, ao meu pai, Valter, pelo seu carinho e incentivo nos momentos difíceis e ao meu irmão Fábio por sua grande amizade.

À Emanuele, minha namorada e melhor amiga, por tudo de bom que ela representa na minha vida. Pela sua compreensão e incentivo a seguir adiante.

À minha orientadora Prof. Dra. Alaine M. Guimarães por confiar em mim, por ter me proporcionado todas as condições para o desenvolvimento deste trabalho e por ter, com seu exemplo, me ensinado a ser um professor e pesquisador melhor.

Ao meu co-orientador Prof. Eduardo Fávero Caires por ter me cedido suas bases de dados e por prontamente me atender sempre que foi necessário.

Ao Prof. Dr. José Carlos Ferreira da Rocha, a quem muito admiro por sua simplicidade e inteligência. Suas explicações durante estes dois anos foram fundamentais.

Aos demais professores do programa com os quais tive contato, pois todos contribuíram de alguma maneira para este trabalho.

Ao Instituto Federal Catarinense por ter me permitido participar deste Programa de Pós-Graduação.

RESUMO

A interessabilidade de regras é uma área da mineração de dados que tem por objetivo reduzir a quantidade de modelos a serem analisados por especialistas na etapa de interpretação do conhecimento descoberto em bases de dados. Embora existam várias medidas de interesse de regras voltadas para as tarefas de associação e classificação, observa-se uma falta de métodos consolidados para análise de interessabilidade de modelos de regressão.

Neste trabalho foi desenvolvido um método para analisar a interessabilidade de modelos de regressão, o qual combina um filtro baseado em modelos probabilísticos multivariados com filtros baseados em testes estatísticos de correlação de Pearson e de postos de sinais de Wilcoxon, resultando em uma nova medida de interessabilidade denominada Impacto.

O método desenvolvido foi aplicado sobre modelos de regressão encontrados no processo de mineração de dados para estimativa de gesso agrícola. Estes dados resultam de três experimentos sob Sistema Plantio Direto realizados na Região dos Campos Gerais, PR, nos quais foram medidos, em diferentes épocas, os teores dos nutrientes do solo após a aplicação de doses de gesso.

Os resultados mostraram que o filtro probabilístico multivariado foi capaz de filtrar os melhores modelos segundo uma visão de utilidade, ou seja, de potencial de aplicação agronômica. Foram selecionados seis modelos com score de Impacto > 0,5, ou seja, considerados interessantes, e destes apenas um foi considerado incorretamente classificado. Por outro lado, os filtros baseados em testes estatísticos foram capazes de filtrar seis modelos sendo que dois deles podem ser considerados incorretamente classificados.

Os atributos identificados como mais relevantes para o problema do gesso agrícola foram a época, o teor de Ca e a concentração de Ca em relação à capacidade de troca catiônica efetiva (CTCe), especialmente em camadas superficiais do solo.

Palavras-chave: gesso agrícola, redes bayesianas, Sensitividade, Impacto

ABSTRACT

The interestingness area of data mining process aiming to reduce the amount of models to be analyzed for experts in the interpretation step of the knowledge discovery in databases.

In this work, a method for analysis the interestingness of regression models was developed. This method combine probabilistic multivariate models with Pearson correlation test and Wilcoxon signed-rank test resulting in a new interestingness measure, named Impact.

The developed method was applied over regression models found during a data mining process for estimating agricultural gypsum requirements.

The results showed that the probabilistic multivariate filter was able to filter the best models according to a utility-based approach, in this case, for practical application on agriculture. Six models were considered interesting, with Impact score > 0.5, and only one was miscategorized. On the other hand, the combined statistical test filters were able to filter six models two of them were miscategorized.

The attributes identified as most relevant to estimate gypsum rate were: time, Ca and its concentration on effective cation exchange capacity (CaCTCe), mainly in superficial layers.

Keywords: gypsum, bayesian networks, Sensitivity, Impact

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS

ACP	Análise de Componentes Principais
ACPS	Análise de Componentes Principais Supervisionada
AM	Aprendizado de Máquina
API	Application Programming Interface
ARFF	Attribute-Relation File Format
BIRCH	Balanced Iterative Reducing and Clustering Using Hierarchies
CTCe	Capacidade de Troca Catiônica efetiva
CSV	Comma-separated values
CURE	Clustering Using Representatives
DBSCAN	Density Based Spatial Clustering of Applications with Noise
DMA	Data Mining Analysis
Eclat	Equivalence Class Transformation
EM	Expectation Maximization
EMA	Erro Médio Absoluto
GAO	Grafo Acíclico Orientado
IFEMiner	Interestingness Filtering Engine
KDD	Knowledge Discovery in Databases
MaxEnt	Máxima Entropia
RB	Rede Bayesiana
UEPG	Universidade Estadual de Ponta Grossa
Weka	Waikato Environment Knowledge Analysis
MVC	Modelo-Visão-Controlador
XML	Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

A descoberta de conhecimento em bases de dados (KDD – do inglês *Knowledge Discovery in Databases*) é uma atividade que envolve a mineração de dados e possui aplicações nas mais diversas áreas, como por exemplo, estimativa de vendas, análise de bolsas de valores, descoberta de novos corpos celestes, procura de conteúdo na web, mineração de comportamento de usuários em web sites, identificação de padrões em imagens médicas e na investigação sobre o relacionamento de variáveis agronômicas.

O KDD é um processo interativo e iterativo envolvendo os seguintes passos: (i) desenvolver um entendimento do domínio da aplicação; (ii) criar uma base de dados para estudo; (iii) efetuar a limpeza e o pré-processamento; (iv) aplicar redução de dados e projeção; (v) escolher a tarefa de mineração de dados: associação, classificação, regressão ou agrupamento; (vi) escolher o algoritmo de mineração a ser aplicado; (vii) efetuar a mineração de dados; (viii) interpretar os padrões minerados – com participação decisiva de um especialista[1]; e, (ix) consolidar o conhecimento descoberto (FAYYAD, 1996; NING, KUMAR e STEINBACH, 2013).

Muitos dos padrões encontrados na etapa (vii) são considerados redundantes, evidentes ou contraditórios entre si ou com a teoria estabelecida (SILBERSCHATZ; TUZHILIN, 1996). Considerando isso, surgiu uma subárea da mineração de dados denominada análise de interessabilidade de regras, cujo objetivo é desenvolver técnicas para identificar as regras mais interessantes e assim reduzir o trabalho do especialista na etapa (viii).

A interessabilidade, ou seja, a capacidade de uma regra ser interessante, não possui um conceito único. Ela pode ser definida considerando que um padrão é interessante se ele sugere uma mudança em um modelo estabelecido (MASOOD; SOONG, 2013), ou ainda, como estando intimamente relacionada com: concisão, cobertura, confiabilidade, peculiaridade, diversidade, novidade, surpresa, utilidade ou acionabilidade de uma regra ou padrão (GENG; HAMILTON, 2006).

A interessabilidade foi desenvolvida originalmente no contexto da tarefa de análise associativa e rapidamente foi vinculada também à análise de classificação. Nos trabalhos de Massod e Soong (2013), Geng e Hamilton (2006), McGarry (2005), Lenca et al. (2008), Wu, Chen e Han (2010) e Guillaume, Grissa e Nguifo (2012), por exemplo, foram discutidas várias medidas e abordagens para análise de interessabilidade, entre elas: suporte, confiança,

[1] Neste texto um especialista é definido como alguém com amplo conhecimento na área de estudo.

interesse, variância, Gini, Kullback, ganho, correlação, Piatetsky-Shapiro, inesperabilidade, acionabilidade, entre muitas outras. Contudo, tanto as medidas quanto as abordagens foram desenvolvidas/aplicadas nas tarefas de associação e classificação.

A motivação deste trabalho foram os resultados de Silva (2013), nos quais bases de dados sobre aplicação de gesso agrícola foram submetidas ao processo de mineração de dados, com o objetivo de encontrar modelos numéricos que representassem a estimativa de aplicação de gesso em função do comportamento dos atributos do solo. Sobre as bases de dados foram feitos diversos tratamentos nas etapas (iii) e (iv) do KDD relacionadas ao pré-processamento, incluindo, por exemplo, uso de técnicas para seleção de atributos e ampliação de registros na base de dados. Posteriormente foi aplicado o algoritmo *M5Rules* (HOLMES; HALL; FRANK, 1999) que realiza a tarefa de regressão e, finalmente, na etapa de pós-processamento (viii), as regras geradas com as execuções do algoritmo foram analisadas com auxílio de um especialista. Essa última etapa demandou um tempo e conhecimento considerável do especialista porque havia 67 modelos de regressão. Tornar esta análise tão mais automatizada quanto possível é uma contribuição significativa e possibilitaria, por exemplo, analisar novos experimentos sobre a predição da dose de aplicação de gesso agrícola de forma mais ágil.

Considerando isso, o objetivo geral deste trabalho foi discutir a interessabilidade de regras de regressão em mineração de dados agrícolas de modo a filtrar os melhores modelos para predição da dose de gesso em função de atributos químicos do solo segundo uma visão de aplicabilidade/utilidade.

Com isso, foram delineados os seguintes objetivos específicos:

- Explorar diferentes critérios, objetivos e subjetivos, de interesse de regras visando construir uma ferramenta capaz de explorar o conhecimento prévio do especialista para encontrar as regras de regressão mais interessantes.
- Modelar e implementar um ambiente para análise de interessabilidade de modelos de regressão compreendendo:
  - interpretadores para os modelos resultantes da execução de algoritmos de regressão do software Weka (Weka 3: Data Mining Software in Java, 2015) para permitir comparações de modelos gerados por diferentes algoritmos.
  - interface gráfica única que permita executar diferentes filtros de interessabilidade;

  - configuração simplificada para os filtros disponibilizados de modo a facilitar a repetição de experimentos;
  - padrão de saída de resultados para permitir a comparação de resultados de diferentes filtros.
- Estabelecer uma nova medida de interesse de regras de regressão.
- Desempenhar a análise de interessabilidade sobre os modelos resultantes do trabalho de Silva (2013).

As próximas seções estão organizadas da seguinte maneira: na seção 2 é feita uma revisão bibliográfica onde é abordada a mineração de dados, especialmente a análise de regressão e o algoritmo *M5Rules*, a interessabilidade de regras e, por fim, a aplicação de gesso agrícola; na seção 3 é apresentado um algoritmo proposto para analisar a interessabilidade de modelos de regressão, o RegFilter, bem como uma nova medida de interessabilidade denominada *Impacto*; na seção 4 discute-se a análise de interessabilidade de modelos de regressão obtidos por Silva (2013) utilizando o método da seção 3; na seção 5 apresenta os resultados e discussão; na seção 6 são apresentadas as considerações finais e na seção 7 as indicações de trabalhos futuros.

2 REVISÃO DE LITERATURA

Nesta seção serão discutidas as tarefas de mineração de dados; alguns conceitos, métricas e abordagens para interessabilidade de regras, o algoritmo *M5Rules* do qual foram extraídas as regras que passaram pela análise de interessabilidade na seção 4 e, por fim, uma fundamentação sobre aplicação de gesso agrícola.

2.1 Mineração de Dados

A mineração de dados envolve o uso de técnicas de aprendizado de máquina (AM) que podem ser utilizadas para predizer ou descrever comportamentos a partir de uma base de dados. Como exemplo de métodos preditivos pode-se citar a classificação e a regressão. Por outro lado, agrupamentos e regras de associação são classificados como métodos descritivos (NING; KUMAR; STEINBACH, 2013).

As regras de associação são utilizadas para procurar relacionamentos entre elementos de uma base de dados normalmente composta por transações, tais como vendas de um supermercado, sessões de um *web site*, entre outras. Os principais algoritmos nesta tarefa são o *Apriori* (AGRAWAL; IMIELINSKI; SWANI, 1993), *FP-Growth* (HAN; PEI; YIN, 2000), *Partition* (SAVASERE; OMIECINSKI; NAVATHE, 1995) e *Eclat* (ZAKI; PARTHASARATHY; LI, 1997) e suas respectivas variantes (WITTEN; FRANK; HALL, 2011).

O resultado da aplicação de algoritmos de associações são regras na forma $A \rightarrow B$ onde A e B representam conjuntos de itens e são independentes entre si. Um exemplo de regra de associação pode ser:

$$\text{salário} = \text{médio}, \text{tempo no emprego} = \text{médio} \rightarrow \text{adimplente}$$

As medidas de avaliação das regras de associação mais utilizadas são o suporte e a confiança, definidas no quadro 1, porém, como discutido nos trabalhos de Geng e Hamilton (2006), Guillaume, Grissa e Nguifo (2012) e Masood e Soong (2013), há mais de meia centena de medidas que podem ser aplicadas a esta tarefa.

O agrupamento consiste em descobrir grupos de dados com alto grau de similaridade intra-grupos, ou seja, entre os elementos do grupo, e baixo grau de similaridade intergrupos. As medidas geralmente utilizadas para determinar a similaridade entre os dados são a distância euclidiana, caso os atributos sejam contínuos, e a distância Manhattan para atributos discretos (NING; KUMAR; STEINBACH, 2013). Como exemplos de algoritmos de agrupamentos têm-se: *BIRCH* (ZHANG; RAMAKRISHNAN; LIVNY, 1996), *DBSCAN*

(ESTER *et al.*, 1996), *CURE* (GUHA; RASTOGI; SHIM, 1998) e *k-Means* (MACQUEEN, 1967).

A tarefa de classificação tem por objetivo encontrar uma função capaz de rotular itens de um banco de dados em um conjunto de classes. Pode ser utilizada em uma abordagem descritiva, na qual se procura por um modelo capaz de resumir informações contidas num banco de dados, ou em uma abordagem preditiva, na qual se aplica o modelo sobre o conjunto de dados para tentar rotular automaticamente um registro que não se conheça a classe. Existem algoritmos classificadores baseados em árvores de decisão: *ID3* (QUINLAN, 1986), *C4.5* e *J48* (QUINLAN, 1993); em regras: *JRIP* (COHEN, 1995) e *PART* (FRANK; WITTEN, 1998); redes neurais: *Multilayer Perceptron* (WITTEN; FRANK; HALL, 2011); máquinas de vetor de suporte: *LibSVM* (CHANG; LIN, 2014), *SMOreg* (SHEVADE *et al.*, 2000) e bayesianos: *Naive Bayes* (JOHN; LANGLEY, 1995), entre outros (NING; KUMAR; STEINBACH, 2013).

Os resultados dos algoritmos de classificação costumam ser apresentados como um conjunto de regras no formato $A \rightarrow B$ ou em modelos que possam ser transformados para este formato. Como exemplo de regra tem-se:

Se salário > R$ 3.500,00 e idade > 26 então risco de inadimplência = médio

Por outro lado, a regressão é uma técnica de modelagem preditiva onde a variável alvo a ser avaliada é contínua (NING; KUMAR; STEINBACH, 2013). Constitui-se de uma metodologia estatística que utiliza o relacionamento entre duas ou mais variáveis quantitativas de modo que a variável meta possa ser predita a partir de outras sob a forma de uma representação funcional (KUTNER *et al.*, 2004).

A escolha da forma funcional da relação de regressão está ligada às variáveis preditoras. Em alguns casos, uma teoria relevante pode indicar a forma funcional apropriada. Em geral, funções de regressão lineares ($f = ax + b$) ou quadráticas ($f = ax^2 + bx + c$) apresentam aproximações satisfatórias. O resultado da análise de regressão é uma relação estatística e não exatamente funcional. Afirma-se isso porque ela não é perfeita. Normalmente as observações para uma relação estatística não se encaixam diretamente na curva de relacionamento, ou curva de regressão (KUTNER *et al.*, 2004).

Para desempenhar uma análise de regressão é necessário, assim como em outras tarefas de mineração de dados, selecionar cuidadosamente as variáveis preditoras. Algumas considerações importantes são: o grau com que cada observação das variáveis pode ser

obtida mais acuradamente, rapidamente ou economicamente quando comparado com outras variáveis; e o grau em que a variável pode ser controlada (KUTNER *et al.*, 2004).

Além da forma funcional e das variáveis, é necessário selecionar cuidadosamente a técnica de regressão que será adotada. Existem vários algoritmos que podem ser utilizados para desempenhar regressão e eles apresentam resultados diferentes com os mesmos dados de entrada. O software *Weka* (WITTEN; FRANK; HALL, 2011) possui, por padrão, os algoritmos: *Gaussian Processes* (MACKAY, 1998); *LeastMedSq* (ROUSSEEUW; LEROY, 2005); *PaceRegression* (WANG, 2000); *SMOReg* (SHEVADE *et al.*, 2000), *M5P* (QUINLAN, 1992); *M5Rules* (HOLMES; HALL; FRANK, 1999); *Isotonic Regression*, *LinearRegression*, *SimpleLinearRegression*, *MultilayerPerceptron* e *RBFNetwork* (WITTEN; FRANK; HALL, 2011).

O algoritmo *M5Rules*, por exemplo, tem por objetivo explorar os bons resultados de modelos de árvores de decisão ao predizerem atributos numéricos, associando-os à simplicidade da apresentação de resultados em listas de regras *IF ... THEN*. Em suma, o algoritmo consiste em um mecanismo para induzir listas de decisão com equações a partir de árvores (HOLMES; HALL; FRANK, 1999).

Primeiramente, o *M5Rules* cria uma árvore de decisão podada sobre toda a base de dados de treinamento. A cada nodo folha é associado um modelo linear sendo que o nodo com o melhor modelo, segundo alguma heurística, é transformado em uma regra. Todas as instâncias cobertas pela regra, ou seja, aquelas contidas na ramificação da árvore correspondente ao modelo linear escolhido são removidas da base de dados de treinamento e o processo é repetido até que todas as instâncias estejam cobertas por alguma regra (HOLMES; HALL; FRANK, 1999).

De acordo com (SILVA, 2013), o algoritmo *M5Rules* já foi aplicado em diferentes áreas de estudo, como exemplo pode-se citar: previsão de resultados de ensaios de fluência em aços; modelagem da estimativa de esforço de projetos de software; estimativa da tenacidade *Charpy* de soldas metálicas; estimativa da força compressiva do concreto; estimativa de valores de temperatura máxima, mínima e média no contexto da agrometeorologia. Guimarães *et al*. (2015) utilizaram o algoritmo *M5Rules* para estimar as doses de gesso agrícola empregadas em sistema plantio direto[2] encontrando que a saturação de Ca na capacidade de troca de cátions efetiva (CTCe) foi um atributo mais importante do que a saturação por Al para estimativa do gesso.

[2] Sistema de manejo caracterizado pelo não revolvimento do solo.

A análise de interessabilidade sobre regras de regressão, objeto de estudo deste trabalho, foi aplicada sobre os resultados de Silva (2013) que são modelos de regressão para a dose de gesso agrícola obtidos com o algoritmo *M5Rules*. Por isso, as subseções 2.2 e 2.3 detalham a interessabilidade de regras e a aplicação de gesso agrícola, respectivamente.

2.2 Interessabilidade de Regras

A interessabilidade é uma subárea da mineração de dados que tem por objetivo estudar mecanismos para avaliar os padrões minerados sob determinados critérios de interesse. Geng e Hamilton (2006) fizeram um extenso levantamento de medidas e abordagens para análise de interessabilidade de regras das quais derivaram nove critérios para afirmar se um padrão é interessante.

- Concisão: um padrão é conciso se ele contém relativamente poucos pares de atributos-valor enquanto que um conjunto de padrões é conciso se ele contém relativamente poucos padrões. No contexto da regressão usando o *M5Rules* significa ter modelos com poucas regras e essas poucas regras serem formadas por poucos atributos.
- Abrangência: um padrão é geral se ele cobre relativamente um grande subconjunto de uma base de dados. Se um padrão caracteriza muita informação de uma base de dados então ele tende a ser mais interessante. O suporte é uma medida que exemplifica o critério de abrangência, visto que regras com alto valor de suporte representam associações com elevada freqüência na base de dados. Algum cuidado é necessário nesta avaliação, pois regras com elevada ocorrência podem representar conhecimento óbvio ou redundante.
- Confiabilidade: um padrão é confiável se o relacionamento descrito pelo padrão ocorrer com alta porcentagem de casos aplicáveis. Por exemplo, uma regra de classificação é confiável se suas predições são altamente acuradas. Em regressão corresponde a uma análise sobre as medidas de erro (médio absoluto, médio relativo, quadrático).
- Peculiaridade: um padrão é peculiar se ele for muito distante de outros padrões descobertos segundo alguma medida de distância. Padrões peculiares costumam ser gerados a partir de dados peculiares, também denominados de *outliers*. São interessantes porque podem representar conhecimento novo para o especialista. No contexto da regressão isso poderia ocorrer com a

localização de modelos que trabalhem com variáveis consideradas de pouca influência sobre o problema em estudo.

- Diversidade: um padrão é diverso se ele contém elementos que diferem significativamente dos demais. Por exemplo, em sumarização, um padrão pode ser considerado diverso se sua distribuição de probabilidade se distanciar de uma distribuição uniforme.
- Novidade: um padrão é novo para alguém se este alguém não o conhecia antes ou não poderia inferi-lo a partir de outros padrões conhecidos. Como não é possível representar todo o conhecimento de um especialista, a novidade não pode ser medida explicitamente em relação à ignorância de um especialista. Porém, podem ser considerados novos, em relação ao conhecimento representado, aqueles padrões que não poderiam ser inferidos a partir dele.
- Inesperabilidade ou surpresa: um padrão é surpreendente se ele entrar em contradição com o conhecimento ou expectativa de uma pessoa. Padrões surpreendentes são interessantes porque podem identificar falhas no conhecimento prévio. A surpresa difere da novidade porque a novidade é um padrão novo e não contradiz qualquer padrão anteriormente conhecido ou esperado.
- Utilidade: um padrão é útil se ele puder contribuir para chegar a um objetivo. Como os objetivos podem variar de pessoa para pessoa, a utilidade é um critério subjetivo e de difícil mensuração. Em análise de regressão, mesmo regras que sejam capazes de predizer uma variável com grande confiabilidade podem ser inúteis. É fundamental observar a importância de cada um dos atributos envolvidos.
- Acionabilidade: um padrão é acionável se, em algum domínio, ele habilitar-nos a tomar decisões sobre ações futuras. A acionabilidade, assim como a surpresa, é de difícil mensuração.

Estes nove critérios podem ser classificados em três grupos: objetivos, subjetivos e semânticos. Os critérios objetivos são baseados apenas nos dados e em medidas estatísticas, probabilísticas ou extraídas da teoria da informação. As medidas subjetivas levam em consideração tanto os dados quanto o conhecimento prévio do domínio, seja ele obtido por consulta a um especialista ou extraído da bibliografia. Por fim, as medidas semânticas

consideram o significado e a apresentação (estrutura) dos padrões. Alguns pesquisadores consideram este um tipo especial de medida subjetiva uma vez que é necessário expressar o conhecimento sobre o domínio de alguma forma (FREITAS, 1999; SILBERSCHATZ e TUZHILIN, 1996; GENG e HAMILTON, 2006; YAO, CHEN e YANG, 2006).

O quadro 1 apresenta as principais medidas objetivas de interessabilidade segundo Freitas (1999), Lenca et. al. (2008) e Geng e Hamilton (2006). Entretanto, existem inúmeras medidas que não estão no quadro, mas que podem ser consultadas nos trabalhos de Guillaume, Grissa e Nguifo (2012), García e Bueno (2012), Glass (2013), entre outros.

É importante destacar que as medidas de interessabilidade discutidas são aplicáveis a regras de associação e classificação e que, possivelmente poderiam ser estendidas para a regressão (NING; KUMAR; STEINBACH, 2013).

Quadro 1. Principais medidas objetivas de interessabilidade

Medida	Fórmula[3]	Descrição
Suporte	$P(AB)$	Segundo Agrawal, Imielinski e Swani (1993) o suporte de uma regra é uma medida de sua significância estatística, ou seja, do percentual de vezes que o antecedente e o conseqüente aparecem juntos na base de dados.
Confiança / Precisão	$P(A \mid B)$	A confiança mede a força de uma regra, ou seja, qual é a possibilidade de obter A sendo que temos B. Sua equação é equivalente à probabilidade condicional (AGRAWAL; IMIELINSKI; SWANI, 1993).
Acurácia	$P(AB) + P(\neg A \neg B)$	É uma medida que reflete a probabilidade dos elementos de A e B ocorrerem juntos ou não ocorrerem. São fatores redutores da acurácia as ocorrências de um dos elementos isolados. Isso mostra que não estão fortemente ligados. Uma regra com acurácia máxima (1) seria tal que os elementos só ocorrem na base de dados se estiverem juntos.
Interesse (lift)	$\frac{P(B \mid A)}{P(B)}$ ou $\frac{P(AB)}{P(A)P(B)}$	De acordo com Brin *et al.* (1997) o interesse é essencialmente uma medida de independência. No entanto, ela só mede a co-ocorrência e não a implicação, uma vez que é simétrica.
Jaccard	$\frac{P(AB)}{P(A)} + P(B) - P(AB)$	Compara o número de presenças comuns e o número de presenças totais excluindo o número de ausências conjuntas. Ou seja, mede a semelhança entre conjuntos de amostras (NING; KUMAR; STEINBACH, 2013).
Convicção	$\frac{P(A)P(\neg B)}{P(A\neg B)}$	A convicção é normalizada sobre o antecedente e o conseqüente como uma noção estatística de correlação. É uma medida de implicação porque é direcional (BRIN *et al.*, 1997). Se a probabilidade de ter A e não ter B simultaneamente aumentar, significa que a convicção sobre a regra diminui. Ao mesmo passo, se a probabilidade de A ou a probabilidade de não B aumentar, também aumenta a convicção. Isso evidencia o caráter direcional e não simultâneo.

[3]Para todas as medidas A representa o antecedente e B o conseqüente de padrões na forma $A \rightarrow B$

P (Coeficiente de Correlação de Person)	$\frac{P(AB) - P(A)P(B)}{\sqrt{P(A)P(B)P(\neg A)P(\neg B)}}$	Mede o grau de uma correlação e a direção da correlação, ou seja, se é positiva ou negativa. O valor 1 significa correlação perfeita positiva, o valor -1 indica correlação perfeita negativa e o valor zero indica que não há correlação entre o antecedente e o conseqüente (PEARSON, 1986),(NING; KUMAR; STEINBACH, 2013).
Ganho de Informação[4]	$\log \frac{P(AB)}{P(A)P(B)}$	O ganho de informação compara a probabilidade de observar A e B juntos com a probabilidade de observá-los isoladamente. Se houver uma associação genuína entre A e B, a P(AB) será >> P(A)P(B) e o ganho >> 0. Por outro lado, se a associação for fraca P(AB) será próximo a P(A)P(B) e o ganho será próximo de zero (CHURCH; HANKS, 1990).

Fonte: Autoria própria.

Silberschatz e Tuzhilin (1996) propuseram medidas subjetivas que levam em consideração o conhecimento do especialista associado aos dados. Elas são adequadas quando:

- O contexto do conhecimento do usuário varia, ou seja, é possível surgirem novos conhecimentos sobre o domínio;
- O interesse do usuário varia e,
- O conhecimento do usuário evolui, ou seja, o novo conhecimento deve ser acrescentado ao banco de conhecimento.

Para verificar o interesse por uma regra usando medidas subjetivas é necessário avaliar o grau de inesperabilidade e de acionabilidade. Em geral, ambos os critérios costumam ocorrer juntos, ou seja, é possível concentrar a busca por regras inesperadas porque elas têm alta probabilidade de serem também acionáveis.

Para estudar o problema os autores consideraram um sistema de crenças contendo restrições na forma *IF .. THEN*. Tais crenças refletem o conhecimento do especialista e podem ser classificadas em *hard beliefs* e *soft beliefs*. As primeiras são crenças que não podem ser alteradas por novos conhecimentos. Quando uma regra contradiz tal crença assume-se que a regra está errada. Podem ser, por exemplo, conhecimentos amplamente comprovados em trabalhos científicos e sobre os quais se acredita não existirem mais dúvidas. Por outro lado, as *soft beliefs* são crenças que podem ser revisadas ou mesmo refutadas se novos conhecimentos forem adicionados por meio do processo de mineração. Elas podem ser, por exemplo, conhecimentos que ainda não foram amplamente analisados em trabalhos, ou conhecimentos que ainda não estão totalmente formulados.

[4] O símbolo log na equação do ganho refere-se ao logaritmo de base 2.

As medidas subjetivas de interessabilidade podem ser aplicadas sob diferentes abordagens, como destacam Guillaume, Grissa e Nguifo (2012):

> **Abordagem 1**: Especifica-se o conhecimento sob alguma linguagem formal de representação, depois se efetua o processamento e, então, o sistema filtra padrões inesperados para exibi-los ao usuário. Em Bellandi *et al*. (2007), por exemplo, é sugerido o uso da linguagem OWL para representação do conhecimento sobre o domínio; Mansing, Bryson e Reichgelt (2011) também empregaram ontologias para fazer esta representação. Por outro lado, Jaroszewics e Simovici (2004); Chan (2006); Jaroszewics (2009) e Malhas e Aghbari (2009) utilizaram redes bayesianas para esta tarefa. Com relação ao ranqueamento das regras, existem ainda trabalhos que utilizaram a máxima entropia (MaxEnt), como Bie (2011) e Bie, Kontonasios e Spyropoulou (2010).
>
> **Abordagem 2:** O usuário interage com o sistema informando o que é interessante ou não, assim, removem-se regras não interessantes. É a abordagem mais comum, mas também a mais morosa e cansativa.
>
> **Abordagem 3:** O sistema aplica as especificações do usuário como restrições durante o processo de mineração para reduzir o espaço de busca e fornecer menos resultados. Isso é utilizado, por exemplo, pelos algoritmos *Apriori* e em alguns algoritmos de árvore como o *C4.5* e o *J48*.

Este trabalho foi baseado na primeira abordagem conforme discutido na seção 3.

2.2.1 Sensitividade de redes bayesianas para avaliar a interessabilidade de regras

Uma rede bayesiana (RB) pode ser definida como um grafo acíclico orientado (GAO) em que cada nó é identificado com informações de probabilidade sendo que (RUSSEL; NORVIG, 2013):

1. Cada nó corresponde a uma variável aleatória, que pode ser discreta ou contínua;

2. Um conjunto de vínculos orientados ou setas conecta pares de nós. Se houver uma seta do nó X até o nó Y, X será denominado pai de Y.

3. Cada nó X tem uma distribuição de probabilidade condicional $P(X \mid pais(X))$ que quantifica o efeito dos pais sobre o nó.

Redes Bayesianas tem sido utilizadas para análise de interessabilidade como uma ferramenta para representação do conhecimento do especialista. Jaroszewicz e Simovici (2004) reduziram o número de itens freqüentes em análise associativa baseados na diferença

absoluta entre o suporte estimado sobre os dados e o suporte obtido a partir da rede bayesiana com resultados promissores, em outro trabalho Jaroszewiecs, Scheffer e Simovici (2008) também utilizaram RB para representar as probabilidades e os relacionamentos de variáveis segundo o conhecimento de um especialista. Foi definido um algoritmo para analisar a interessabilidade comparando o comportamento entre os dados observados e a rede bayesiana. Malhas e Aghbari (2009) propuseram uma abordagem subjetiva para verificar a interessabilidade de conjuntos de atributos, utilizando uma medida que eles denominaram de Sensitividade calculada pelo algoritmo *Interestingness Filtering Engine* (*IFEMiner*), sobre RB.

O *IFEMiner* toma como entradas uma RB representando o conhecimento do domínio, a qual tem seus valores de probabilidade calculados a partir de uma base de casos; a própria base de casos; um valor mínimo de suporte e um valor mínimo para o *threshold*[5]. O algoritmo busca estimar a inesperabilidade de um *conjunto de atributos interessantes*[6] utilizando o incremento do potencial de incerteza de uma RB quando um padrão é atribuído como descoberta na rede. Em outras palavras, o *score* calculado pelo *IFEMiner* representa o quanto um padrão dado como entrada numa rede aumenta a incerteza sobre as crenças pré-estabelecidas.

A crença em uma RB é a probabilidade posterior, condicionada sobre todas as descobertas atualmente atribuídas à RB. A informação mútua de Shannon (SHANNON, 1948), que é baseada na entropia, é uma medida do quanto de informação pode ser obtida sobre uma variável aleatória pela observação de outra variável e, portanto, pode ser utilizada para calcular a Sensitividade.

Considerando uma base de dados com um conjunto N de variáveis categóricas ou discretas e uma RB contendo os mesmos N nós, pode-se afirmar que, de acordo com Pearl (1988), a incerteza residual no valor de uma variável meta X, dado que Y é instanciado para um valor y, pode ser calculada como:

[7]

$$H(X \mid y) = -\sum_{x} P(x, y) \log P(x, y) \qquad (1)$$

[5] Também conhecido como limiar, refere-se a um valor mínimo de alguma quantidade. Neste caso o valor mínimo para a Sensitividade de um conjunto de atributos interessante.

[6] O *IFEMiner* considera todos os conjuntos de atributos encontrados com suporte e threshold mínimos maiores do que os valores estipulados pelo especialista como conjuntos de atributos interessantes.

[7] O símbolo log nesta equação significa o logarítimo de base 2.

A incerteza residual média em X, somada sobre todas as possíveis saídas da variável Y é dada por.

$$H(X \mid Y) = \sum P(x,y) \log \frac{P(y)}{P(x,y)} \tag{2}$$

E, portanto, o potencial de redução da incerteza total de Y, ou seja, sua informação mútua, M, em relação a X é dada por:

$$M(X;Y) = H(X) - H(X \mid Y) \tag{3}$$

Se RB é uma rede bayesiana sobre um conjunto de atributos N, (I,i) é um conjunto de itens obtido com o algoritmo *Apriori* executado sobre a base de dados, tal que $I \subseteq N$ e $i \in Dom(I)$. E, seja também Q um nodo *query* tal que $Q \in N$, e F seja um nodo *finding* tal que $F \in N$. A sensitividade S de um conjunto de itens (I,i) com relação à RB é definida como:

$$S(I,i) = \sum_{m=1}^{N} \sum_{n=1}^{N} M_{new}(Q_m;F_n) = \sum_{m=1}^{N} \sum_{n=1}^{N} H_{new}(Q_m) - H_{new}(Q_m \mid F_n) \tag{4}$$

tal que M_{new} só é incrementado ao valor de $S(I,i)$ se $M_{old}(Q_m;F_n) < M_{new}(Q_m;F_n)$, onde $M_{old}(Q_m;F_n)$ é a informação mútua entre dois nodos quaisquer antes das entradas de novas descobertas e $M_{new}(Q_m;F_n)$ após.

Por fim, definida a sensitividade de um conjunto de itens, é possível encontrar a sensitividade de um conjunto de atributos utilizando a equação 5.

$$S(I) = \max_{i \in Dom(I)} S(I,i) \tag{5}$$

Um resumo do algoritmo *IFEMiner* está no quadro 2 e o algoritmo completo encontra-se em (MALHAS; AGHBARI, 2009).

O suporte dado como entrada para o *IFEMiner* é utilizado na execução do *Apriori* para reduzir os conjuntos de itens apenas àqueles que possuírem freqüência maior do que um valor especificado pelo usuário. De modo similar o *threshold* é utilizado para reduzir a quantidade de conjuntos de atributos na medida em que são removidos da listagem aqueles que possuírem valor inferior ao informado pelo usuário.

Quadro 2. Algoritmo *IFEMiner*

Algoritmo:*IFEMiner*

Entrada: Uma rede bayesiana RB representando o conhecimento sobre o domínio, suporte mínimo, um valor mínimo para o threshold e uma base de dados.

Saída: Conjuntos de atributos interessantes ordenados pelo seu score de sensitividade.

1. Encontrar os conjuntos de itens freqüentes (I,i) na base de dados usando o algoritmo *Apriori* (AGRAWAL; IMIELINSKI; SWANI, 1993);

2. Computar a sensitividade dos itens freqüentes S(I,i) (Equação 4);

3. Computar a sensitividade dos conjuntos de atributos S(I) (Equação 5);

4. Retornar os conjuntos de atributos interessantes resultantes do passo 3 com valor de sensitividade maior do que o threshold.

Fonte: (MALHAS; AGHBARI, 2009)

2.2.2 Testes estatísticos para avaliar o interesse em modelos de regressão

2.2.2.1 Teste estatístico sobre a correlação de uma população

Uma forma de verificar a qualidade do ajuste de um modelo de regressão é calcular o coeficiente de determinação R^2 que pode variar de 0 a 1. Seu valor próximo de 1 significa que a maioria da variabilidade observada na variável resposta pode ser expressa pelo modelo de regressão. O coeficiente de determinação pode ser obtido elevando o coeficiente de correlação ao quadrado (NING; KUMAR; STEINBACH, 2013).

O coeficiente de correlação de Pearson é uma medida estatística que representa a força e a direção de uma relação **linear** entre duas variáveis. O Símbolo *r* representa o coeficiente de correlação amostral, determinado pela equação 6 (LARSON; FARBER, 2009).

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}} \tag{6}$$

Assim como o coeficiente de determinação, o coeficiente de correlação também pode variar de 0 a 1. Se *x* e *y* apresentam uma correlação linear positiva forte, *r* está próximo de 1. Se, por outro lado, *x* e *y* têm uma correlação linear negativa forte, *r* está próximo de -1. Quando não há correlação linear ou esta é fraca, então *r* está próximo de 0. Isso não significa, contudo, que não há relação entre *x* e *y*, apenas que esta não é linear (LARSON; FARBER, 2009).

É possível utilizar um teste de hipóteses para determinar se o coeficiente de correlação de uma amostra *r* fornece evidência suficiente para concluir que o coeficiente de correlação de uma população ρ é significante (KUTNER *et al.*, 2004). Neste caso as hipóteses seriam:

$$H_0 : \rho = 0 \text{ (não há correlação significante)}$$
$$H_a : \rho \neq 0 \text{ (há correlação significante)}$$

Um teste *t* pode ser usado para decidir se H_0 deve ser rejeitada ou se o correto é falhar em rejeitá-la. A estatística de teste é dada pela equação 7 que segue uma distribuição *t* com n – 2 graus de liberdade.

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \tag{7}$$

Como exemplo de aplicação pode-se considerar os dados da tabela 1 que representam os valores preditos por dois diferentes modelos de regressão.

Tabela 1. Valores preditos para a variável y por dois modelos de regressão distintos.

y	y (Modelo 1)	y (Modelo 2)
2,0	2,4	3,0
3,0	2,8	3,4
4,0	3,2	3,8
5,0	4,0	5,9
6,0	6,5	5,5
7,0	7,1	4,9

Fonte: autoria própria.

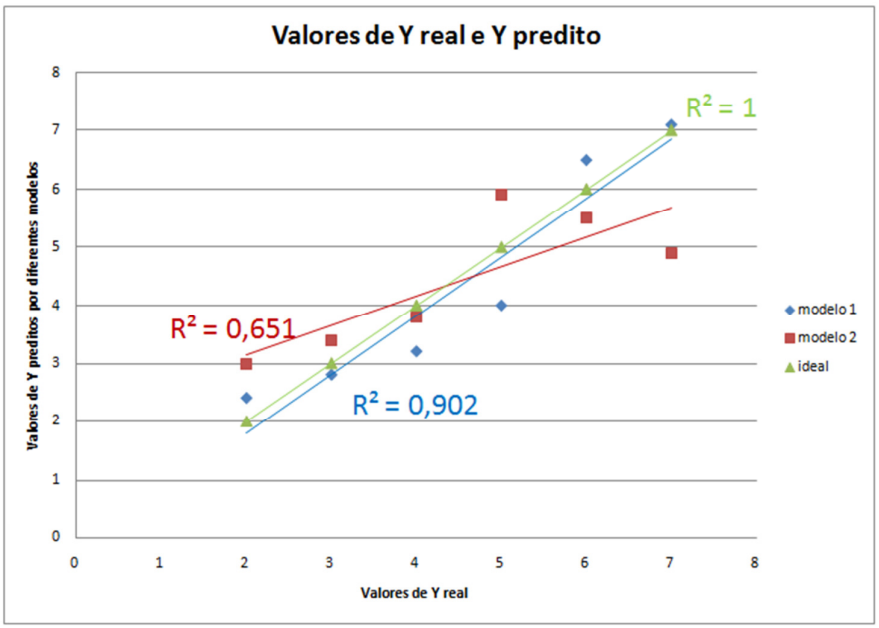
O gráfico da figura 1 ilustra as regressões lineares para cada um dos modelos e aquela que seria a ideal. A tabela 2 apresenta os valores calculados para R^2, *r* e *t*.

Tabela 2. Valores de coeficiente de determinação, coeficiente de correlação e estatística t para os modelos 1 e 2

	Modelo 1	Modelo 2
Coeficiente de determinação	0,902	0,651
Coeficiente de correlação	0,95	0,81
Estatística t	6,07	2,73

Fonte: autoria própria.

Figura 1. Dispersão de pontos para os valores de Y real e Y predito segundo os modelos 1 e 2



Fonte: autoria própria.

Considerando um teste *t* bi-caudal[8] com nível de significância $\alpha = 0{,}05$, e n – 2 = 4 graus de liberdade os pontos críticos para rejeição seriam $t < -2{,}776$ e $t > 2{,}776$.

Como $t_{\text{modelo1}} > 2{,}776$, então se deve rejeitar H_0, ou seja, ao nível de significância de 5%, há evidência suficiente para concluir que há correlação linear significativa entre os valores reais e os valores preditos pelo modelo 1. Por outro lado, como $-2{,}776 < t_{\text{modelo2}} < 2{,}776$, deve-se falhar em rejeitar H_0, ou seja, ao nível de significância de 5% conclui-se que não há uma correlação linear significativa entre os valores reais e os valores preditos pelo modelo 2.

Diante do exposto, o teste de hipótese para o coeficiente de correlação linear de uma população pode ser utilizado como um filtro de interessabilidade. O interesse seria por modelos capazes de apresentar correlação linear significativa entre os valores preditos e os valores reais. Isso é especialmente interessante em casos onde se tenha dados para testes diferentes daqueles utilizados na construção do modelo. É possível verificar se o modelo é útil para generalizações, ou seja, se poderia ser aplicado a outras bases de dados de forma segura.

[8] Neste caso um teste t bi-caudal é utilizado porque o coeficiente de correlação admite valores entre -1 e 1 e, como mostrado pela equação 7, valores negativos para r implicariam em estatísticas de teste t também negativas.

2.2.2.1 Teste estatístico de Postos de Sinais de Wilcoxon

A um conjunto de dados pode ser associada uma distribuição de probabilidade. Em alguns casos tal distribuição pode ser aproximada a alguma distribuição bem conhecida, chamadas de distribuições paramétricas. Um dos métodos mais empregados para estimar os valores dos parâmetros destas distribuições é o Método da Máxima Verossimilhança. Contudo, um grande obstáculo à utilização prática do método consiste na freqüente incapacidade de se obter uma solução explícita para a maioria dos problemas. Nestes casos é necessário utilizar algum método de otimização numérica para obtenção dos parâmetros de interesse (PORTUGAL, 2014).

Sob o ponto de vista da interessabilidade, poderia ser útil saber se os valores preditos para uma variável *y* provêm de uma população com a mesma distribuição de probabilidade que os dados reais de *y*.

Num cenário em que não se conhece a distribuição de probabilidade da população é possível utilizar o teste de postos de sinais de Wilcoxon. Trata-se de um teste não paramétrico que pode ser utilizado para determinar se duas amostras dependentes foram selecionadas de populações que possuem a mesma distribuição (LARSON; FARBER, 2009).

Neste teste as hipóteses seriam:

H_0 : Não há diferença entre os valores preditos e os valores reais.

H_a : Há diferença entre os valores preditos e os valores reais.

Para exemplificar, considerando o conjunto de dados da tabela 3, teríamos:

Tabela 3. Desenvolvimento do teste de postos de sinais de Wilcoxon para valores hipotéticos de y real e y predito

y real	y predito	Diferença	Valor absoluto	Posto	Postos de Sinais
2,1	2,4	-0,3	0,3	6	-6
1,8	1,6	0,2	0,2	4	4
2,8	2,6	0,2	0,2	4	4
4,5	5,0	-0,5	0,5	9	-9
3,7	3,5	0,2	0,2	4	4
4,0	4,1	-0,1	0,1	1,5	-1,5
5,5	6,0	-0,5	0,5	9	-9
4,7	5,1	-0,4	0,4	7	-7
5,2	5,3	-0,1	0,1	1,5	-1,5
2,5	3,0	-0,5	0,5	9	-9

Fonte: autoria própria.

A soma dos postos negativos é -43 e a soma dos postos positivos é 12. Com isso, a estatística de teste $w_s = 12$. Consultando uma tabela para o teste de postos de sinais de Wilcoxon para n = 10 e $\alpha = 0{,}05$ em um teste bi-caudal verifica-se que $w = 8$. Como $w_s > w$ deve-se decidir falhar em rejeitar a hipótese nula, ou seja, conclui-se ao nível de significância de 5% que não há diferença entre os valores preditos e os valores reais.

Diante do exposto, o teste de postos de sinais pode ser utilizado como um filtro de interessabilidade de modo a selecionar somente modelos para os quais os valores preditos apresentem a mesma distribuição de probabilidade que os valores reais, mesmo sem saber qual é a forma da distribuição.

2.3 Gesso agrícola

A acidez do solo limita a produção agrícola em consideráveis áreas no mundo, em decorrência da toxidez causada por Al e Mn e pela baixa saturação por bases. A calagem consegue obter efeitos de redução de Al e Mn tóxicos além de fornecer Ca e Mg. Contudo, sua ação é, geralmente, limitada ao local de aplicação no solo, que em Sistema Plantio Direto, é restrita à superfície (COLEMAN e THOMAS, 1967; CAIRES *et al*., 2003).

Como os materiais corretivos da acidez utilizados na agricultura são pouco solúveis, os produtos da reação do calcário com o solo têm mobilidade limitada. A correção do subsolo ácido poderia ser feita por meio da calagem profunda, todavia, essa prática necessita de revolvimento do solo, e por isso não é de interesse em áreas já estabelecidas com o sistema plantio direto. Além disso, a calagem profunda exige máquinas potentes e equipamentos caros, o que torna a prática onerosa (CAIRES *et al.*, 1998).

Uma alternativa viável é a aplicação do gesso agrícola. Isso se deve ao fato do gesso ser relativamente solúvel e, quando aplicado na superfície do solo movimentar-se ao longo do perfil com a influência do excesso de umidade. Ao atingir o subsolo, o gesso aumenta o suprimento de Ca^{2+} e reduz a toxidade do Al^{3+}. Como resultado dessa melhoria do subsolo, as raízes são capazes de se desenvolver em maior profundidade, permitindo maior eficiência na absorção de água e nutrientes (CAIRES *et al*., 1998, 1999).

O gesso agrícola ($CaSO_4.2H_2O$) é um subproduto das indústrias de fertilizantes fosfatados. Na sua composição química básica, o gesso agrícola contém os elementos cálcio (17 a 20%) e água livre (15 a 20%). A aplicação de 1t ha^{-1} de gesso com 15% de umidade adiciona aproximadamente 200 kg de cálcio, 160 kg de enxofre e 8 kg de fósforo na forma de P_2O_5. A quantidade de cálcio adicionada, neste caso, eleva o seu teor na camada de solo

entre 0 e 20 cm em cerca de 0,5 $cmol_c\ kg^{-1}$. Na solução do solo o gesso dissocia-se pela hidrólise segundo a equação 8 (NUERNBERG; RECH; BASSO, 2005):

$$CaSO_4.2H_2O \rightarrow Ca^{2+} + SO_4^{2-} + CaSO_4^0 \quad (8)$$

Os íons Ca^{2+} e SO_4^{2-}, após a dissociação, participam do complexo de troca de cátions e ânions. O par iônico $CaSO_4^0$ move-se no perfil do solo, facilitando a descida de complexos químicos solúveis neutros ($CaSO_4^0$, $MgSO_4^0$ e KSO_4^0) ao subsolo. A intensidade desta movimentação depende da composição mineral e orgânica, da textura e, principalmente da estrutura e das condições climáticas (PAVAN, 1982).

2.3.1 Recomendações de Gesso Agrícola

As recomendações da necessidade de gesso agrícola existentes são consideradas conservadoras, pois levam 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[9]. Ao mesmo tempo, 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 (Raij *apud* Silva, 2013).

Para os solos da região dos cerrados a recomendação da necessidade da dose de gesso é feita com base no teor de argila do solo, sendo a equação 9 para culturas anuais e a 10 para culturas perenes.

$$N.G(kg/ha) = 50 \times Argila(\%) \quad (9)$$

$$N.G(kg/ha) = 75 \times Argila(\%) \quad (10)$$

Raij (2008) afirma que ao utilizar o teor de argila do solo para quantificação da dose de gesso perde-se sustentação das equações para solos com propriedades eletroquímicas e de capacidade de retenção diferentes. A tabela 4, apresenta recomendações de gesso com base na textura do solo.

[9] Neste contexto é um processo de transporte de nutrientes (Mg, K, etc) rumo às camadas mais profundas do solo.

Tabela 4. Recomendação de gesso agrícola (15% S) em função da classificação textural do solo para culturas perenes e anuais.

Textura do Solo	Dose de gesso agrícola ($kg.ha^{-1}$)	
	Culturas anuais	Culturas perenes
Arenosa	700	1050
Média	1200	1800
Argilosa	2200	3300
Muito argilosa	3200	4800

Fonte: (SOUSA; LOBATO; REIN, 2005)

Ribeiro *et al.* (1999) apresentaram a recomendação de gesso agrícola do estado de Minas Gerais, que é feita com base nos teores de Ca^{2+} e Al^{3+}. Quando os teores de Ca^{2+} forem iguais ou inferiores a 4 $mmol_c.dm^{-3}$ e/ou Al^{3+} for maior do que 5 $mmol_c.dm^{-3}$ e/ou quando a saturação por Al é maior do que 30% nas camadas 20 a 40 cm ou 30 a 60 cm devem ser utilizados os valores da tabela 5.

Tabela 5. Recomendação de gesso baseado no teor de argila segundo a recomendação do Estado de Minas Gerais.

Argila ($g.kg^{-1}$)	N.G ($t.ha^{-1}$)
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

Fonte: (RIBEIRO; GUIMARÃES; ALVAREZ, 1999)

Alvarez *et al. apud* Melo e Silva (2006) sugerem que a quantidade de gesso a ser utilizada seja 25% da necessidade de calagem calculada tanto pelo método do Al^{3+} trocável, quanto pela saturação por bases. O valor obtido é corrigido em relação à profundidade multiplicando-se a necessidade de gesso pelo quociente entre a camada que se quer corrigir (em cm) e 20 cm que é a camada para a qual foi calculada a calagem.

Resumindo, os métodos em uso no Brasil para estimar a necessidade de gesso (NG, em kg ha^{-1}) visando amelhoriado ambiente radicular no subsolo foram baseados no teor de argila (g kg^{-1}) nas camadas subsuperficiais: NG = 5 x argila (SOUSA; LOBATO, 2002) e NG = 6 x argila (RAIJ *et al.*, 1996). Embora o teor de argila seja um atributo do solo que pode ser usado na estimativa da dose de gesso, tais métodos são incompletos (RAIJ, 2008) e não estimam com precisão as doses mais econômicas de gesso a serem empregadas. Há outros métodos sugeridos para estimar a necessidade de gesso (VITTI *et al.*, 2008) que ainda carecem de comprovação científica quanto à sua eficiência.

2.3.2 Uso da mineração de dados para extração de conhecimento agronômico envolvendo o uso de gesso agrícola

Silva (2013) explorou diferentes abordagens para obtenção de modelos de predição da necessidade de gesso agrícola como função dos atributos químicos do solo. Para isso, foram utilizados dados oriundos de três locais na região dos Campos Gerais, Paraná.

A primeira área, denominada A, está situada no município de Ponta Grossa (25º 8' S, 50º 15´ W a uma altitude de 853 m), em Latossolo Vermelho de textura média (295 g kg^{-1} de argila) e alta acidez. No momento em que foi instalado o experimento, a área se encontrava sob sistema de plantio direto há 15 anos. Na área foram aplicadas doses de gesso agrícola (0, 4, 8 e 12 t ha^{-1}) à lanço na superfície em 1993. Foi utilizado delineamento de blocos completos ao acaso com três repetições (CAIRES *et al*., 1998, 1999).

A segunda área, denominada B, também está localizada no município de Ponta Grossa (25º 10' S, 50º 05' W e altitude média de 970 m), em um Latossolo Vermelho de textura argilosa (580 g kg^{-1} de argila) com acidez média. Foram aplicadas doses de gesso agrícola (0, 3, 6 e 9 t ha^{-1}) à lanço sobre a superfície do solo no ano de 1998 respeitando um delineamento de blocos completos ao acaso com três repetições (CAIRES, FELDHAUS, BLUM, 2001; CAIRES et al., 2002, 2003).

Por fim, a terceira área, denominada C, está localizada no município de Guarapuava (25º 17' S, 50º 05' W, em altitude média de 1080 m). Trata-se de um Latossolo Vermelho de textura argilosa (650 g kg^{-1} de argila) com acidez média. Foram aplicadas doses de gesso (0, 4, 8, 12 t ha^{-1}) à lanço sobre a superfície do solo em 2005 respeitando um delineamento de blocos completos ao acaso com quatro repetições (CAIRES *et al.*, 2011).

Em todas as áreas foram feitas coletas em quatro níveis de camadas do solo: A (0-10 cm), B (10-20 cm), C (20-40 cm) e D (40-60 cm). Variaram entre os experimentos as épocas das coletas. Na região A as coletas ocorreram aos 8, 20, 32 e 44 meses após a aplicação do gesso; na região B as coletas ocorreram aos 8, 20, 32, 44, 56, 68, 80 e 92 meses após aplicação do gesso e, na região C as coletas foram aos 8 e 20 meses após a aplicação. O clima das três áreas experimentais é classificado como Cfb segundo o sistema Köppen-Geiger, com verões frescos e geadas freqüentes durante o inverno. A temperatura média anual é de 17º C. A precipitação pluvial média anual é de 1600 mm, com chuvas concentradas entre os meses de setembro e abril.

Para todas as áreas foram determinados os teores de Al, Ca, Mg e K ($mmol_c$ dm^{-3}) trocáveis utilizando os métodos estabelecidos pelo Instituto Agronômico do Paraná

(PAVAN *et al.*, 1992). A capacidade de troca catiônica efetiva (CTCe) foi determinada com a equação 11.

$$CTC_e = Al + Ca + Mg + K \tag{11}$$

A saturação por Al, Ca, Mg e K na CTCe foi calculada pela equação 12 onde X pode ser substituído pelo teor de Al, Ca, Mg ou K.

$$\text{Saturação por } X(\%) = (X \,/\, CTCe) \times 100 \tag{12}$$

O resultado final foi uma base de dados com 42 atributos considerando as segmentações nos 4 níveis do solo (A, B, C e D): Al (A,B,C,D), Ca (A,B,C,D), Mg (A,B,C,D), K (A,B,C,D), m (A,B,C,D), CaCTCe (A,B,C,D), Mg (A,B,C,D), K (A,B,C,D), época e gesso. Para todos os atributos químicos do solo foram determinados os valores de AB, dados pela média dos valores das camadas A e B, o que gerou mais 8 atributos.

Foram feitas três separações das bases de dados:

- BD1 contendo os dados das áreas A, B e C, com 8 e 20 meses após a aplicação de gesso totalizando 106 registros.
- BD2 contendo os dados das áreas A e B com 8, 20, 32 e 44 meses após a aplicação de gesso. Total de 144 registros.
- BD3 contendo os dados da área B, com 8, 20, 32, 44, 56, 68, 80 e 92 meses após a aplicação de gesso, totalizando 132 registros.

Também foram trabalhadas subdivisões da base de dados e as abordagens de Análise de Componentes Principais (ACP) e Análise de Componentes Principais Supervisionada (ACPS) para seleção de atributos. Como resultados da execução do *M5Rules* foram obtidos 67 modelos distintos, os quais se encontram no anexo A juntamente com o valor do coeficiente de correlação e do erro médio absoluto calculados pelo *Weka*. No trabalho de Silva (2013) foi discutido um conjunto de 18 regras, para as quais o quadro 3 mostra um resumo das análises. Tais análises foram conduzidas com auxílio de um especialista em gesso agrícola. O número à frente da equação representa o número de ordem do modelo tal como aparece no anexo A.

Quadro 3. Análise de Silva (2013) para os modelos obtidos com o M5Rules.

Modelo	Resumo das análises	
50-1: gesso = 0,0959 * epoca + 1,0021 * Al_AB – 0,2187 * Ca_C + 0,1141 * Mg_AB + 0,2193 * Mg_C + 1,7824 * K_C + 0,3183 * Ca_CTCe_AB + 0,2188 * Ca_CTCe_C – 0,5856 * K_CTCe_C – 29,3984	O modelo mescla atributos de teor e de saturação, o que o torna agronomicamente inviável. Além disso, ele contém muitos atributos, são 9, o que o torna muito específico e de baixa cobertura. Nota-se ainda a relação inversamente proporcional entre gesso e o teor de Ca, o que é incoerente uma vez que o gesso é uma fonte de cálcio (OLIVEIRA; PAVAN, 1996), (CAIRES; FELDHAUS; BLUM, 2001), (MARQUES, 2008); (NEIS *et al.*, 2010). Como a saturação é obtida a partir do teor de cátions, se ambos estiverem presentes no modelo eles devem ter a mesma correlação (positiva ou negativa) com o atributo estimado.	IA[10] VM
51-2: gesso = 0,0888 * epoca + 0,1482 * Ca_CTCe_AB – 0,2998 * Mg_CTCe_AB + 2,3871	O modelo 51-2 tem coerência com os trabalhos de Oliveira e Pavan (1996); Caires *et. al*. (1998, 2004) e Soratto e Crusciol (2008), em que a aplicação do gesso aumentou os teores de Ca e diminuiu os teores de Mg nas camadas superficiais do solo.	VA VM
52-3: gesso = 0,1675 * Ca_CTCe_AB – 0,2647 * Mg_CTCe_AB + 1,4675	O modelo é coerente pela mesma justificativa da regra 51-2.	VA VM
53-4: gesso = 0,0664 * epoca + 0,2748 * Ca_CTCe_AB – 12,714	Ao verificar a consolidação do modelo foi observado que houve diferença entre a dose estimada e a quantidade real aplicada de gesso inferior a 1 t/há, um resultado muito bom.	VA VM
54-5: gesso = +0,179 * Ca_C + 0,2375 * Mg_C – 0,1599 * Mg_CTCe_C + 4,257	O modelo 5 é contraditório, nele 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).	IA IM
55-6: gesso = 0,0396 * epoca + 0,2891 * Ca_AB – 0,2809 * Mg_AB – 1,8504	O modelo 55-6 está coerente com os trabalhos de Oliveira e Pavan (1996); Caires *et. al*. (1998, 2004) e Soratto e Crusciol (2008), nos quais a aplicação de gesso aumentou o teor de Ca e diminuiu o teor de Mg na camada superficial do solo (0-20). Na consolidação do modelo, na maioria dos casos, houve boa estimativa da dose de gesso, sendo a diferença menor que 1 tonelada por hectare.	VA VM
56-7: gesso = 0,032 * epoca + 0,2842 * Ca_AB – 0,2613 * Mg_AB – 0,5218 * K_AB + 0,1077	O modelo 7 apresenta boa aplicabilidade. Dentre 18 registros testados, apenas dois apresentaram diferenças entre a dose estimada e a dose aplicada superior a 2 toneladas por hectare. Apesar da 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). Portanto, além do teor de Mg o atributo teor de K também apresentou correlação negativa com o atributo gesso.	VA VM
57-8: gesso = 0,1168 * epoca + 0,0051 * Ca_CTCe_AB – 0,3459 * Mg_CTCe_AB – 0,3584 * K_CTCe_AB + 9,7584	Na consolidação deste modelo houve grande diferença entre as doses estimadas e a quantidade real de gesso. Uma possível explicação foi a presença do atributo $KCTCe_{AB}$ pois a lixiviação do K não ocorreu em todas as áreas de estudo. Sendo assim, não é interessante buscar modelos que utilizem o potássio pois o mesmo não apresenta um comportamento regular em relação à lixiviação.	IA VM
58-9: gesso = 0,0545 * epoca + 0,137 * Ca_CTCe_AB – 0,2296 * Mg_CTCe_AB + 1,6416	Os atributos que compõe o modelo 58-9 são os mesmos do modelo 51-2 e, portanto, valem as mesmas considerações.	VA VM

[10] IA = inviável agronomicamente; VA = viável agronomicamente; IM = inviável matematicamente e VM = viável matematicamente.

59-10: gesso = 0,051 * epoca + 0,2423 * Ca_CTCe_C – 7,688	O modelo 10, apesar de apresentar ótimas respostas para a estimativa de gesso, apresentou valores de r (0,63) e EMA (2,50) considerados não satisfatórios computacionalmente. Porém, na discussão final, este modelo foi inserido entre os mais interessantes.	VA IM
60-11: gesso = 0,0677 * epoca + 0,1575 * Ca_CTCe_AB – 0,2815 * Mg_CTCe_AB + 1,5963	Foi utilizada a técnica SMOTE para gerar este modelo. Ele possui as mesmas características dos modelos 58-9 e 51-2.	VA VM
61-12: gesso = 0,0727 * epoca + 0,1494 * Ca_CTCe_AB – 0,2925 * Mg_CTCe_AB + 2,3616	Também utilizou a técnica SMOTE. Foram selecionados os atributos época, Mg_{AB}, $CaCTCe_{AB}$ e $MgCTCe_{AB}$ a partir do método ACPS, contudo, os atributos utilizados foram os mesmos dos modelos 60-11, 58-9 e 51-2.	VA VM
62-13: gesso = 0,0485 * epoca + 0,2579 * Ca_A – 0,1783 * Mg_B – 5,3371	O modelo 62-13 apresentou ótimas estimativas para a dose de gesso. Em sete casos a diferença entre a dose estimada e a quantidade real aplicada de gesso foi menor que 0,8 toneladas por hectare.	VA VM
63-14: gesso = 0,108 * epoca – 0,2967 * Mg_CTCe_A – 0,1204 * Mg_CTCe_B – 0,3339 * K_CTCe_C + 15,7802	Foram utilizados todos os atributos referentes à saturação de todas as camadas (A,B,C,D), mais os atributos época e gesso. O que dificulta a aplicabilidade dos modelos foi o fato de envolverem atributos de todas as camadas do solo.	IA VM
64-15: gesso = 0,1186 * epoca + 0,2663 * Ca_CTCe_A + 0,0788 * Ca_CTCe_C – 0,043 * Mg_CTCe_D- 17,5739	Eles apresentaram coerência por conterem os atributos saturação de Mg e K em relação inversamente proporcional à dose de gesso aplicada. Embora sejam agronomicamente coerentes, são de difícil aplicação por envolverem o mesmo atributo em diferentes camadas.	IA VM
65-16: gesso = 0,0202 * epoca + 0,2174 * Ca_CTCe_A + 0,0503 * Ca_CTCe_D – 12,1985	O modelo 16 foi considerado adequado para a estimativa de gesso. O interessante deste modelo é o fato dele envolver apenas os atributos época, saturação por Ca na camada mais superficial (0-10) e saturação por Ca na camada mais profunda (40-60 cm).	VA VM
66-17: gesso = 0,119 x Ca_CTCe_C – 0,2687 x Mg_CTCe_A – 0,1288 x K_CTCe_D + 5,8379	O modelo 17 pode ser considerado agronomicamente coerente pois a saturação por Ca foi diretamente proporcional ao gesso, e a saturação por Mg e K inversamente proporcional à dose de gesso. Há coerência com os trabalhos de Quaggio *et al*. (1993), Caires *et. al.* (1998) e Rampin (2008). Contudo, o modelo apresenta atributos de diferentes camadas de solo (A, C e D), dificultando a sua aplicação para a estimativa de gesso.	VA VM
67-18: gesso = 0,121 * Ca_A + 0,3446 * Ca_D – 0,121 * Mg_C – 2,6925	O modelo 67-18 foi obtido utilizando dados selecionados com a matriz de covariância, assumindo a independência entre os atributos das camadas A e D e pode ser considerada agronomicamente coerente. Apresentou ótimas estimativas, contendo nove casos em que a diferença entre a dose estimada e a quantidade real aplicada foi igual ou menor do que 1 tonelada por hectare. Porém, o fato de serem teores de camadas diferentes dificulta a aplicação da regra.	VA VM

Fonte: autoria própria

Como resultado final, Silva (2013) concluiu que os melhores modelos para aplicação prática seriam o 53-4, 55-6, 59-10 e 65-16. Todos os dados utilizados e os modelos de regressão produzidos foram disponibilizados para este estudo.

3 MÉTODO DESENVOLVIDO PARA ANALISAR A INTERESSABILIDADE DE MODELOS DE REGRESSÃO

A tarefa de regressão, em geral, acaba por produzir modelos que apresentam somente uma solução numérica para o problema. Selecionar os melhores modelos, segundo uma visão baseada em utilidade ou acionabilidade é uma tarefa que demanda trabalho exaustivo de um especialista no domínio sob estudo. O algoritmo *RegFilter*, proposto neste trabalho, utiliza uma nova medida, denominada Impacto e tem por objetivo encontrar modelos de regressão com bom potencial para aplicação prática. Ele foi incluído no software *Data Mining Analysis* (DMA) descrito brevemente no apêndice A.

3.1 Algoritmo *RegFilter*

O algoritmo *RegFilter* está definido no Quadro 4. Ele utiliza restrições de: tamanho de descrição (número de regras por modelo e número de termos); desempenho do modelo (correlação e erro médio absoluto), atributos mais interessantes (usando o algoritmo *IFEMiner*) e o impacto para apresentar um ranking das regras mais aplicáveis/úteis. São fornecidas como entrada: uma rede bayesiana que modela o conhecimento do especialista, um arquivo descrevendo as correlações entre os atributos descritores e o atributo meta, uma base de dados discretizada, um conjunto de modelos de regressão, o valor mínimo para o suporte (que será utilizado pelo *IFEMiner* em um dos passos do *RegFilter*), os valores mínimos para o threshold e para o coeficiente de correlação de cada modelo bem como os valores máximos para o número de atributos por modelo e para o número de regras por modelo.

Quadro 4. Algoritmo *RegFilter*

Algoritmo: RegFilter

Entradas:

RB /* rede bayesiana representando o conhecimento sobre o domínio; */

CorrInfo /* arquivo contendo o relacionamento de correlação entre as variáveis A1, A2, ... An e a variável meta (estabelecido pela literatura da área ou em consulta a um especialista); */

BD /* base de dados com todos os atributos discretos ou categóricos */

M /* conjunto de k modelos de regressão. */

MinSuport /* suporte mínimo (utilizado pelo *IFEMiner*); */

MinThreshold /* threshold mínimo (utilizado pelo IFEMiner); */

MinCorrCoef /* coeficiente de correlação mínimo; */

MaxMAE /* valor máximo para o erro absoluto médio; */

MaxRules /* número máximo de regras R admitidos no modelo M; */

MaxTerms /* número máximo de termos por regra R admitido no modelo M. */

Saída:

M /* filtrado e ordenado pelo seu score total de Impacto. */

```
1. /* remover modelos com coeficiente de correlação inferior ao MinCorrCoef */
2. filterByCorrelation(M, MinCorrCoef);
3. /* remover os modelos com erro médio absoluto maior do que MaxMAE */
4. filterByMAE(M, MaxMAE);
5. /* remover os modelos com mais regras do que MaxRules */
6. filterByMaxRules(M, MaxRules);
7. /* remover os modelos com mais termos do que MaxTerms */
8. filterByMaxTerms(M, MaxTerms);
9. /* Remover as regras cujos termos não respeitem as restrições impostas por CorrInfo */
10. filterByCorrInfo(M, CorrInfo);
11. /*Encontrar os conjuntos de atributos interessantes segundo o IFEMiner (MALHAS; AGHBARI,
12. 2009). A cada conjunto está associado um score de Sensitividade S*/
13. CAI = IFEMiner(BD, RB, MinSupport, MinThreshold);
14. /* Calcular o score total de Sensitividade de cada modelo conforme método descrito na seção 3.2 */
15. Stotal = calculateModelScores(CAI,M);
16. /* Calcular o score de Sensitividade de cada modelo devido ao conhecimento prévio conforme
17. procedimento descrito na seção 3.2 */
18. Sprior = extractPriorKnowledge(RB, Stotal, M);
19. /* calcular o score de Impacto conforme definição apresentada na seção 3.2 */
20. calculateImpactScore(M, Sprior, Stotal);
21. /* ordenar os modelos pelo valor do score total de Impacto em ordem decrescente */
22. orderByImpact(M);
```

Fonte: autoria própria.

Nas linhas 2 e 4 são removidos os modelos que não atendem a restrições mínimas de desempenho matemático.

Na seqüência, as linhas 6 e 8 aplicam restrições com relação ao tamanho máximo de descrição. Em determinados domínios a utilização de muitas variáveis impossibilita o uso de um modelo. Portanto, o tamanho máximo de descrição pode preservar a utilidade dos modelos embora não afirme qual é o mais útil. Na regressão isso significa controlar o número de termos (variáveis envolvidas) e, se o algoritmo usado for o *M5Rules* ou similares, pode ser controlado também o número de regras por modelo.

A linha 10 contém um procedimento para aplicar restrições quanto ao relacionamento entre as variáveis preditoras e a variável meta. É possível, por exemplo, que inúmeros experimentos revelem que a correlação entre duas variáveis A e B é positiva e isto seja um fato aceito pela comunidade científica. Neste caso, modelos que contenham correlação negativa entre A e B estariam em contradição com o conhecimento científico estabelecido. Conclui-se que, embora possam ser matematicamente interessantes segundo a análise do erro de predição, tais modelos não poderiam ser utilizados por contradizerem a teoria estabelecida. Caso o relacionamento entre as variáveis não seja bem conhecido, ou não seja aceito pela comunidade científica, deve ser possível aceitar modelos que prevejam correlações indeterminadas. Ao se determinar um relacionamento como positivo ou negativo se está determinando *hard beliefs* segundo a definição de Silberschatz e Tuzhilin (1996), ou seja, crenças que não podem ser contraditas. Por outro lado, ao assinalar um relacionamento como indeterminado é estabelecida uma *soft belief*, ou seja, uma crença que pode ser contradita.

Na linha 13 são calculados os conjuntos de atributos interessantes utilizando o algoritmo *IFEMiner* definido em Malhas e Aghbari (2009). Estes conjuntos de atributos interessantes serão posteriormente utilizados para calcular um *score* de Sensitividade total de um modelo regressão, linha 15. O procedimento e a justificativa deste cálculo estão descritos na seção 3.2. O algoritmo *IFEMiner* é utilizado porque ele é capaz de contrastar o relacionamento entre as variáveis estabelecido pelo especialista, com os dados obtidos por via experimental e assim, quantificar o interesse num conjunto de atributos sob a medida Sensitividade. Isso conduz a algum grau de subjetividade na medida em que diferentes especialistas podem entender o domínio de modo diverso e assim correlacionarem as variáveis na rede bayesiana de modo também diferente. Diferentes estruturas de redes bayesianas conduzem a alterações nos conjuntos de atributos interessantes e no valor do interesse em cada conjunto. Em suma, permite verificar o interesse do especialista (utilidade) a partir do seu conhecimento sobre o assunto.

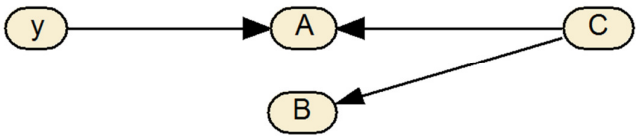
Posteriormente, será calculado, na linha 18, o quanto deste *score* total de Sensitividade se deve a relacionamentos contemplados no corpo de conhecimento prévio, conforme descrito na seção 3.2.

Por fim, na linha 20 é calculado o *score* de Impacto seguindo a definição apresentada na seção 3.2 e, na linha 22, os modelos são ordenados pelo *score* decrescente de Impacto.

3.2 Cálculo do Impacto de um modelo de regressão

É possível calcular o valor do score total de Sensitividade de um modelo de regressão linear, $S_{M\,total}$. Pode-se, por exemplo, considerar o modelo $M : y = aA + bB - cC$, obtido com uma base de dados contendo os atributos *y*, *A*, *B* e *C*. Sobre esta base de dados e sobre a RB da figura 2 é aplicado o algoritmo *IFEMiner* e são encontrados os conjuntos de atributos interessantes $CAI_1 = \{A, B\}$, e $CAI_2 = \{B, C\}$, com Sensitividade $S_{CAI_1} = 2$ e $S_{CAI_2} = 3$.

Figura 2. Rede Bayesiana de Exemplo



Fonte: autoria própria.

Neste caso o modelo *M* seria interessante em um grau 2 devido ao fato de conter os atributos A e B e, também seria interessante em um grau 3 por conter os atributos B e C. Assim, o *score* total de Sensitividade do modelo é $S_{M\,total} = 2 + 3 = 5$.

O *score* total pode ser decomposto como mostrado da equação 13, onde $S_{\mathrm{M}\,total}$ corresponde ao *score* considerando todos os conjuntos de atributos contemplados, $S_{\mathrm{M}\ prévio}$ representa o *score* considerando todos os conjuntos de atributos cujos relacionamentos de dependência estejam previstos na RB e $S_{M\,\neg prévio}$ é o *score* considerando somente os conjuntos de atributos cujos relacionamentos de dependência não estejam previstos na RB.

$$S_{\mathrm{M}\,total} = S_{\mathrm{M}\,prévio} + S_{\mathrm{M}\,\neg prévio} \tag{13}$$

A figura 2 mostra que existe um relacionamento de dependência entre os atributos B e C e que não existe um relacionamento de dependência entre A e B. Portanto, $S_{\mathrm{M}\,prévio} = 3$, $S_{\mathrm{M}\,\neg prévio} = 2$ e $S_{\mathrm{M}\,total} = 5$.

Fazendo uma manipulação algébrica na equação 14 chega-se à definição de uma nova medida, denominada Impacto, dada pela equação 14.

$$I_M = \frac{S_{\mathrm{M}\,prévio}}{S_{\mathrm{M}\,total}} = 1 - \frac{S_{\mathrm{M}\,\neg prévio}}{S_{\mathrm{M}\,total}} \tag{14}$$

Na equação 14 o I_M representa o grau de impacto de um modelo calculado com base no conhecimento prévio, ou seja, desconsiderando conjuntos de atributos que não estejam diretamente relacionados na RB.

O valor de I_M varia entre 0 e 1 sendo que 1 significa que todo o grau de interesse na regra é devido ao impacto do conhecimento prévio, de modo contrário, o valor 0 significa que todo o grau de interesse se deve ao impacto das variáveis que não estavam relacionadas.

Ao se fazer uma análise de regressão deve-se selecionar cuidadosamente as variáveis da base de dados (KUTNER *et al.*, 2004). É muito provável que sejam selecionadas pelo analista as variáveis mais importantes para o domínio de acordo com uma visão de utilidade. Sendo assim, é bem provável que regras que apresentem um $I_M > 0{,}5$ sejam mais interessantes sob o ponto de vista prático (útil) do que regras com $I_M < 0{,}5$. Isto porque é mais provável que variáveis cujo relacionamento seja bem conhecido sejam também mais controláveis, de maior poder de descrição do problema e, assim, apareçam conectadas na rede bayesiana. Relacionamentos de dependência que não se tenha muito conhecimento provavelmente não serão representados por dificuldades na definição dos valores das crenças.

No exemplo, o $I_M = 1 - 2/5 = 0{,}6$. Isso significa que a regra apresenta um caráter maior de utilidade do que de inesperabilidade frente ao corpo de conhecimento apresentado.

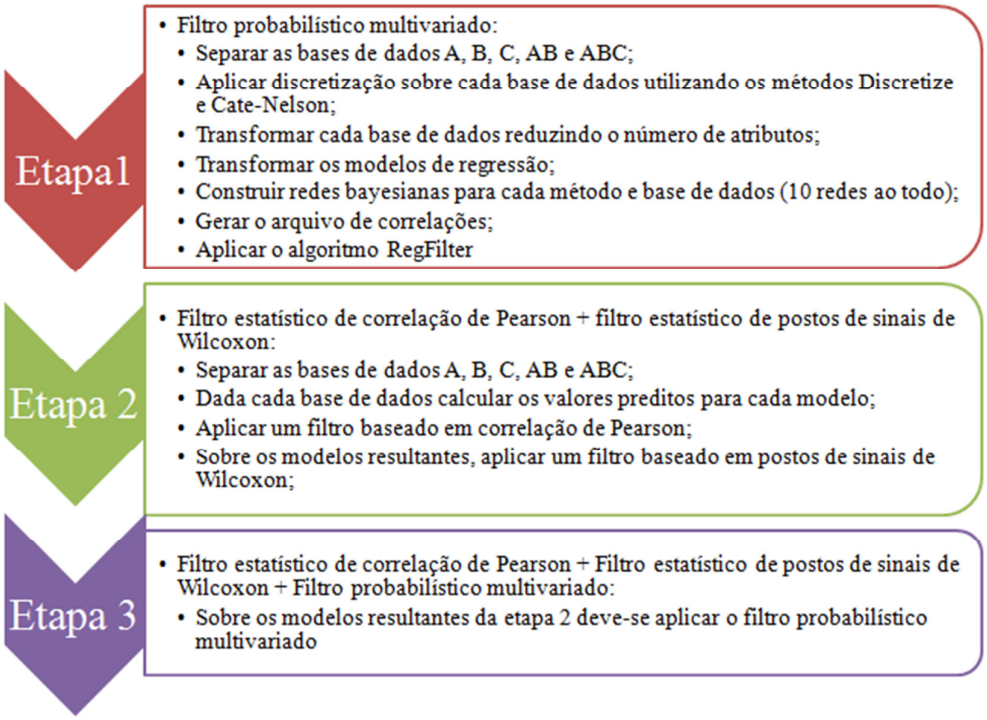
4 ANÁLISE DE INTERESSABILIDADE DE MODELOS DE REGRESSÃO PARA NECESSIDADE DE GESSO AGRÍCOLA

Esta seção discute os procedimentos adotados bem como os resultados obtidos para uma análise de interessabilidade desempenhada sobre modelos de regressão para predição da dose de gesso agrícola utilizando o método proposto neste trabalho, descrito na seção 3.

4.1 Etapas de execução da análise

O objetivo da análise foi investigar a interessabilidade de regras de regressão para encontrar aquelas com maior potencial para aplicação prática. As análises foram divididas em três etapas como ilustra a figura 3.

Figura 3. Representação gráfica das etapas de análise



Fonte: autoria própria.

4.1.1 Etapa 1 – Aplicação do filtro probabilístico multivariado

Nesta etapa o algoritmo *RegFilter* foi aplicado sobre o conjunto de 67 modelos de regressão (Anexo A) variando as bases de dados utilizadas como entrada, ou seja, foram utilizadas as bases das áreas A, B, C, uma junção das bases A e B e, por fim, uma junção das bases A, B e C. Isoladamente e independentemente dos modelos terem sido gerados com

aqueles dados. Tais bases são as mesmas do trabalho de Silva (2013), descritas na seção 2.3.2.

Foi necessário discretizar os dados contínuos e esta tarefa foi feita utilizando duas abordagens distintas. Na primeira foi trabalhada a discretização utilizando três faixas de valores buscando deixá-las balanceadas, ou seja, com o mesmo número de ocorrências em cada faixa. Esta abordagem foi utilizada por ser amplamente utilizada em mineração de dados. Para fazer as discretizações foi utilizado o filtro *Discretize* do *Weka.* Na segunda abordagem foi utilizado o método Cate-Nelson (CATE; NELSON, 1971). Maiores detalhes sobre o método Cate-Nelson podem ser encontrados no Apêndice A.3.

Após a discretização foi necessário reduzir a dimensionalidade da base de dados uma vez que o *NeticaJ*[11], utilizado na implementação do filtro, em sua versão *freeware,* não é capaz de trabalhar com mais de 15 atributos. Além disso, seria necessário projetar uma rede bayesiana com 42 atributos. Dada a restrição no número de instâncias das bases de dados (nenhuma contém mais de 200 registros) e a quantidade elevada de mundos possíveis nas tabelas de probabilidade condicional é possível que nem todos os casos ocorram na base de dados ou que ocorram distorções nos valores das probabilidades e, posteriormente, nos cálculos da Sensitividade e Impacto para cada modelo.

O procedimento adotado foi o seguinte:

1. Foi criado um atributo discreto chamado camada em cada uma das bases de dados. Este atributo pode assumir os valores c1, c2, c3 e c4 correspondendo aos níveis A, B, C e D do solo.
2. Os valores dos atributos Al_A, Ca_A, Mg_A, K_A, m_A, Ca_CTCe_A, Mg_CTCe_A e K_CTCe_A foram associados com a camada c1. Os atributos da camada B com c2 e assim por diante.

As figuras 4 (a) e (b) ilustram o processo de transformação da base de dados. Por simplificação foi considerado somente o atributo Ca nas camadas A e B.

Com os dados discretizados e transformados foram preparadas as redes bayesianas, duas para cada base de dados, sendo uma para a discretização com o *Discretize* e outra para a discretização com o método Cate-Nelson. As probabilidades condicionais foram aprendidas a partir dos arquivos de casos (as próprias bases de dados). As figuras 5 e 6 mostram exemplos de redes bayesianas para as duas abordagens de discretização aplicadas sobre a base de dados ABC.

[11] Foi utilizado a API NeticaJ para trabalhar com as redes bayesianas no software desenvolvido.

Figura 4. (a) base de dados original e (b) base de dados transformada.

gesso	epoca	Ca_A	Ca_B
d0	e1	faixa1	faixa2
d0	e2	faixa2	faixa1
d4	e1	faixa1	faixa2
d4	e2	faixa1	faixa2
d8	e1	faixa2	faixa1
d8	e2	faixa2	faixa1

(a)

gesso	epoca	camada	Ca
d0	e1	c1	faixa1
d0	e2	c1	faixa2
d4	e1	c1	faixa1
d4	e2	c1	faixa1
d8	e1	c1	faixa2
d8	e2	c1	faixa2
d0	e1	c2	faixa2
d0	e2	c2	faixa1
d4	e1	c2	faixa2
d4	e2	c2	faixa2
d8	e1	c2	faixa1
d8	e2	c2	faixa1

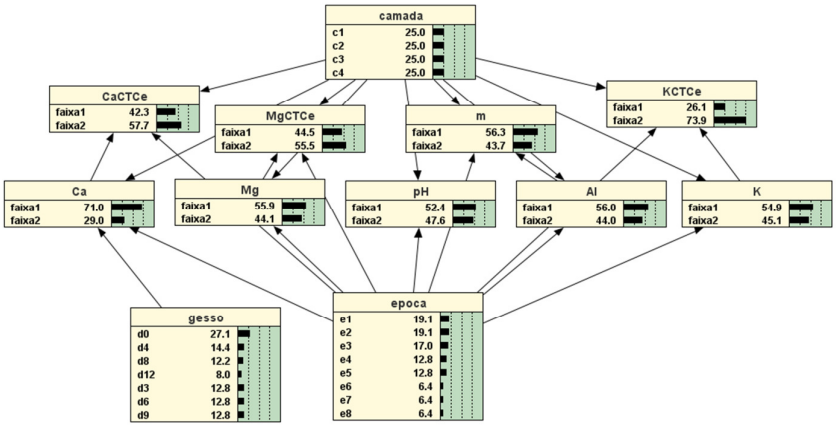
(b)

Fonte: autoria própria.

É importante observar que todas as redes foram construídas para representar os mesmos relacionamentos entre as variáveis. De modo geral, foi definido que a época e a camada condicionam todos os atributos químicos do solo e que o gesso, por ser uma fonte de Ca, condiciona o teor do mesmo. Com isso, um exemplo de entrada na tabela de probabilidades condicionais do Ca seria P(Ca=faixa1 | camada = c3, gesso = d4, epoca = e1) = 0,86.

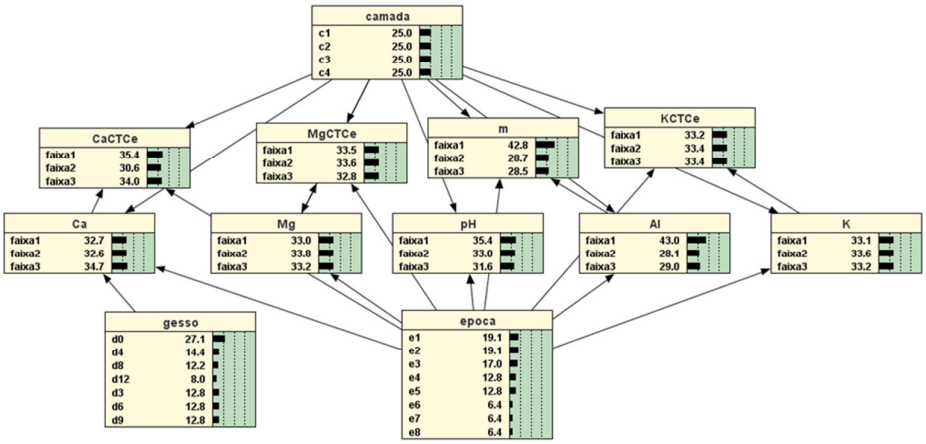
Além disso, também foi modelado que o teor de Ca condiciona o atributo Ca_CTCe; o teor de Al condiciona a saturação por alumínio, m; o teor de Mg condiciona o atributo Mg_CTCe e o teor de K condiciona o atributo K_CTCe.

Figura 5. Rede bayesiana com os atributos discretizados pelo método Cate-Nelson para a base de dados ABC



Fonte: autoria própria.

Figura 6. Rede bayesiana com os atributos discretizados pelo método *Discretize* para a base de dados ABC



Fonte: autoria própria.

Os modelos de regressão foram obtidos por Silva (2013) utilizando a base de dados sem a transformação. Sendo assim, as regras contém atributos como Ca_B e K_C, por exemplo. Por isso foi adotado um procedimento de simplificação no qual as informações sobre as camadas foram suprimidas. A figura 7 (a) mostra o modelo 57-08 do anexo A antes da transformação e a figura 7 (b) mostra o modelo após a transformação.

Figura 7. Modelos de regressão antes (a) e após o processo de transformação (b).

```
gesso =
      +0.1168 * epoca
      + 0.0051 * Ca_CTCe_AB
      - 0.3459 * Mg_CTCe_AB
      - 0.3584 * K_CTCe_AB
      + 9.7584 [48/41.633%]
```

(a)

```
gesso =
      +0.1168 * epoca
      + 0.0051 * CaCTCe
      - 0.3459 * MgCTCe
      - 0.3584 * KCTCe
      + 9.7584 [48/41.633%]
```

(b)

Fonte: autoria própria

As informações sobre a correlação esperada entre o gesso e cada um dos demais atributos foram obtidas em revisão bibliográfica e o resultado está no quadro 5. Nele é possível observar que entre o gesso e o Ca (e o CaCTCe) é esperada uma correlação positiva. Tal afirmação se justifica porque, segundo Oliveira e Pavan (1996) e Caires *et al.* (2001), o gesso é uma fonte de cálcio e a sua aplicação no solo eleva os teores desse nutriente. Sobre a relação do gesso com o pH, sabe-se que o gesso praticamente não influencia o pH (CAIRES *et al*., 1998; QUAGGIO *et al*., 1993; RAIJ, 2008). Por isso, foi indicada uma correlação desconhecida entre eles (nem positiva e nem negativa). Isso não

significa necessariamente que não exista correlação, durante o processamento podem ser aceitos modelos que sugiram tanto correlação positiva quanto negativa.

A correlação entre o gesso e o Mg (e o MgCTCe) foi apontada como negativa. Isso porque o trabalhos de Caires *et al*. (1998), Caires *et al*. (2003), Soratto e Crusciol (2008) e Quaggio *et al*. (1993) mostraram uma redução dos teores de Mg nas camadas superficiais em decorrência da aplicação do gesso. Efeito semelhante ocorre com o K, embora, em sistema plantio direto, em proporção muito inferior. Mesmo assim, o relacionamento entre o gesso e o K (e o KCTCe) também foi sinalizado como negativo.

Por fim, a correlação com o Al e com a saturação por alumínio m foi identificada como negativa. Isto porque os trabalhos de Caires *et al*., (1998), Caires *et al*. (1999), Soratto e Crusciol (2008) e Neis *et al*., (2010) apontam que a aplicação de gesso é capaz de agir sobre a toxidade por Al, reduzindo-a, em alguns casos, pela redução de sua saturação. A relação entre a camada e o gesso não foi analisada porque os modelos não contêm este atributo, ele foi criado no processo de transformação descrito na figura 7.

Quadro 5. Correlação entre o gesso e os demais atributos.

Atributo	Correlação
Ca e CaCTCe	Positiva
Mg e MgCTCe	Negativa
K e KCTCe	Negativa
Al e m	Negativa
pH	Desconhecida

Fonte: autoria própria

Com os dados preparados foram feitas as aplicações do filtro, para as quais os resultados são descritos na seção 5.

4.1.2 Etapa 2 – Aplicação dos filtros estatísticos de correlação de Pearson e postos de sinais de Wilcoxon

Nesta etapa os filtros estatísticos foram aplicados sobre o conjunto de 67 modelos de regressão variando as bases de dados utilizadas como entrada, ou seja, foram utilizadas as bases das áreas A, B, C, uma junção das bases A e B e, por fim, uma junção das bases A, B e C, isoladamente e independentemente dos modelos terem sido gerados com aqueles dados. Esse processo foi semelhante ao utilizado no filtro probabilístico multivariado.

Não foram necessárias transformações, discretizações e nem redução de dimensionalidade nas bases de dados. Contudo, foi necessário calcular o valor da dose de gesso predita para cada um dos modelos.

Primeiramente foi aplicado o filtro estatístico baseado no teste de hipóteses sobre a correlação de uma população. O objetivo foi verificar a existência de uma correlação linear significativa, ao nível de significância $\alpha = 0{,}20$, entre as doses de gesso preditas e as reais. Feito isso, os modelos que apresentaram correlação linear significativa foram submetidos ao teste de postos de sinais de Wilcoxon com o objetivo de verificar se as amostras calculadas e as reais vinham de populações que possuíam a mesma distribuição de probabilidade. Os resultados da aplicação dos dois testes são discutidos na seção 5.

4.1.3 Etapa 3 – Combinação dos resultados

Esta etapa consistiu na aplicação dos filtros anteriores na seguinte ordem: primeiro foi aplicado o filtro estatístico de teste de hipóteses sobre a correlação de uma população; depois foram dados como entrada somente os modelos filtrados com correlação linear significativa para o filtro estatístico de teste de postos de sinais de Wilcoxon. O resultado deste teste foi submetido ao filtro probabilístico multivariado e as melhores regras foram ranqueadas pelo seu valor de Impacto.

Todos os filtros foram aplicados sobre as cinco bases de dados. Cabe destacar que nesta etapa, no filtro probabilístico multivariado foram utilizadas somente as bases de dados discretizados pelo método Cate-Nelson por motivos discutidos na seção seguinte. Os filtros estatísticos utilizaram as mesmas bases de dados da etapa 2.

5 RESULTADOS E DISCUSSÃO

As tabelas 6 e 7 mostram, respectivamente, os valores calculados para o Impacto dos modelos que atenderam às restrições de coeficiente de correlação mínimo de 0,40; suporte mínimo de 0,10; threshold mínimo de 0,00; número máximo de regras por modelo igual a 1 e número máximo de termos por regra igual a 5. Os resultados aparecem ordenados pelo Impacto em ordem decrescente, coeficiente de correlação decrescente e erro médio absoluto crescente.

Foi adotado um valor de suporte que, apesar de baixo, se justifica pelo fato de que os atributos época, gesso e camada possuem valores de freqüência baixos para suas classes. Com isso, se fosse estipulado um valor de suporte de 0,30, por exemplo, nenhuma associação com estes atributos seria encontrada e a eles não seriam atribuídos graus de interesse.

Silva (2013) discutiu somente modelos que continham uma única regra. A justificativa para isso foi a complexidade dos modelos. Alguns modelos gerados continham três, quatro ou até mesmo seis regras, as quais acabavam por serem específicas demais. Isso não é prático sob o ponto de vista agronômico. Por este motivo, a configuração adotada foi de uma regra por modelo. Além disso, na discussão sobre os resultados, resumida no quadro 3, existem restrições ao número de atributos contemplados pelo modelo. Para o número máximo de termos, foi adotado o valor de cinco termos, considerando o termo independente. O valor foi assim definido pelo entendimento de que regras com mais de cinco termos podem ser consideradas automaticamente complexas e de difícil aplicação no domínio da aplicação de gesso agrícola.

Utilizando as bases de dados discretizados com o método Cate-Nelson obteve-se 35 modelos que atenderam às restrições conforme mostra a tabela 6. Fazendo uma análise sobre o desempenho médio em relação à aplicação sobre as cinco bases de dados, 16 modelos possuem Impacto igual a zero, o que significa que não são interessantes sob o ponto de vista da utilidade. Dos 19 modelos com algum grau de interesse, 10 deles apresentam Impacto< 0,5, o que significa que a maior parte do valor do score de Sensitividade do modelo pode ser explicada por associações entre atributos que não estão relacionados pela RB, ou seja, pelo critério de utilidade poderiam ser descartados. Restaram, portanto, as regras rotuladas de 1 a 9, com destaque para os modelos 55-6 (gesso = 0,0396 x época + 0,2891 x Ca_{AB} – 0,2809 x Mg_{AB} – 1,8504) e 62-13 (gesso = 0,0485 x época + 0,25769 x Ca_{A} – 0,1783 x Mg_{B} – 5,3371). Tais modelos envolvem os atributos gesso, época, Ca_{AB} e Mg_{AB} e; gesso, época,

Ca_A, Mg_A, respectivamente. Ambos foram gerados a partir dos dados da base B e tiveram resultados similares em relação ao *r* e EMA.

Ainda sobre o desempenho médio, destacam-se os modelos 53-04 (gesso = 0,0664 x época + 0,2748 x $CaCTCe_{AB}$ – 12,714), 59-10 (gesso = 0,051 x época + 0,2423 x $CaCTCe_C$ – 7,688), 43 (gesso = 0,0312 x época + 0,237 x $CaCTCe_A$ – 11,6847) e 22 (gesso = 0,0211 x época + 0,1917 x $CaCTCe_A$ + 0,0545 x $CaCTCe_D$ – 10,7622). Nota-se que além do gesso e da época, o atributo CaCTCe em diferentes camadas aparece em ambas as regras. Este resultado é coerente com a análise feita por Silva (2013) e por Guimarães *et al*. (2015). Além disso, em Silva (2013) identificou, após longo processo de análise com o especialista, os modelos 53-04, 55-06, 59-10 e 65-16 (gesso = 0,0202 x época + 0,2174 x $CaCTCe_A$ + 0,0503 x $CaCTCe_D$ – 12,1985) como aqueles com maior potencial de aplicação prática. Destes quatro modelos, três foram identificados entre os quatro melhores em média na tabela 6, embora dois deles com CV elevados (> 0,25). Estes modelos, conforme já discutido por Silva (2013) estão coerentes com os resultados experimentais de Caires *et al*., (1998, 2004, 2011), Oliveira e Pavan (1996) e Quaggio *et al*., (1993).

Traçando uma análise sobre o desempenho dos melhores modelos em relação às bases de dados isoladamente, observa-se que, de um modo geral, o desempenho dos modelos foi ruim quando aplicado sobre os dados da base C, a qual contém dados de um solo de textura argilosa e de baixa acidez. Quando foram calculados os *scores* de Impacto sobre os dados da área A, textura média e alta acidez, os modelos 55-6 e 62-13 apresentaram os melhores resultados. Os modelos 53-04, 55-06, 59-10 também foram selecionados, porém apresentaram resultados medianos. Este comportamento se repete em relação aos dados da base AB.

Por fim, cabe ainda destacar que para a região B, solo de textura argilosa e com média acidez, os modelos 53-04, 59-10, 43 e 22 apresentaram o melhor desempenho. Para este tipo de solo é interessante o uso de modelos baseados nos atributos época e CaCTCe, este último em diferentes camadas, mas principalmente na camada superficial A (0-10) e AB (0-20). Para solos de alta ou baixa acidez os modelos necessitam de complementação com um termo envolvendo o Mg, presente nos modelos 55-6 e 62-13, que em conjunto com o Ca apresentou os melhores resultados de aplicabilidade.

Tabela 6. Impacto dos modelos filtrados com o filtro probabilístico multivariado para os dados discretizados com a abordagem de Cate-Nelson.

	Modelos	r	EMA	Cate-Nelson (2 faixas de valores)						
				A	B	C	AB	ABC	I (médio)	CV
1	55-6	0,71	1,86	0,66	0,46	0,49	0,72	0,48	0,56	0,18
2	62-13	0,71	1,90	0,66	0,46	0,49	0,72	0,48	0,56	0,18
3	53-04	0,63	2,42	0,51	1,00	0,33	0,50	0,45	0,56	0,42
4	59-10	0,63	2,50	0,51	1,00	0,33	0,50	0,45	0,56	0,42
5	43	0,62	2,13	0,51	1,00	0,33	0,50	0,45	0,56	0,42
6	22	0,58	2,28	0,51	1,00	0,33	0,50	0,45	0,56	0,42
7	63-14	-	-	0,80	0,39	0,44	0,33	0,44	0,48	0,34
8	61-12	0,82	1,71	0,62	0,43	0,32	0,37	0,51	0,45	0,24
9	51-2	0,82	1,77	0,62	0,43	0,32	0,37	0,51	0,45	0,24
10	56-7	0,72	1,80	0,47	0,46	0,40	0,54	0,36	0,45	0,14
11	14	0,66	2,65	0,45	0,47	0,46	0,47	0,39	0,45	0,07
12	64-15	-	-	0,62	0,43	0,32	0,37	0,51	0,45	0,24
13	58-9	0,66	2,28	0,62	0,43	0,32	0,37	0,51	0,45	0,24
14	60-11	0,62	2,40	0,62	0,43	0,32	0,37	0,51	0,45	0,24
15	35	0,70	2,58	1,00	0,00	0,36	0,00	0,58	0,39	0,97
16	65-16	0,67	1,98	0,51	1,00	0,00	0,00	0,45	0,39	0,95
17	57-8	0,84	1,70	0,51	0,35	0,33	0,34	0,36	0,38	0,18
18	67-18	0,71	1,88	0,34	0,36	0,33	0,49	0,00	0,30	0,53
19	54-5	0,42	2,56	0,34	0,37	0,30	0,42	0,00	0,29	0,52
20	13	0,84	1,79	0,00	0,00	0,00	0,00	0,00	0,00	0,00
21	29	0,82	1,63	0,00	0,00	0,00	0,00	0,00	0,00	0,00
22	66-17	0,82	1,77	0,00	0,00	0,00	0,00	0,00	0,00	0,00
23	52-3	0,81	1,74	0,00	0,00	0,00	0,00	0,00	0,00	0,00
24	48	0,79	1,57	0,00	0,00	0,00	0,00	0,00	0,00	0,00
25	12	0,79	2,07	0,00	0,00	0,00	0,00	0,00	0,00	0,00
26	39	0,78	2,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00
27	33*	0,76	2,48	0,00	0,00	0,00	0,00	0,00	0,00	0,00
28	45*	0,76	2,48	0,00	0,00	0,00	0,00	0,00	0,00	0,00
29	44	0,74	2,46	0,00	0,00	0,00	0,00	0,00	0,00	0,00
30	31	0,74	2,46	0,00	0,00	0,00	0,00	0,00	0,00	0,00
31	37	0,70	2,36	0,00	0,00	0,00	0,00	0,00	0,00	0,00
32	40	0,73	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00
33	46	0,63	2,85	0,00	0,00	0,00	0,00	0,00	0,00	0,00
34	24	0,60	2,15	0,00	0,00	0,00	0,00	0,00	0,00	0,00
35	30	0,53	2,43	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Fonte: autoria própria.

Os resultados da tabela 7 mostram que utilizando a discretização dos dados em três faixas de valores balanceados não foi possível selecionar nenhum modelo utilizando o impacto médio, ao mesmo tempo os valores do coeficiente de variação foram elevados. Numa análise independente sobre o resultado de cada base de dados, tem-se que: todos os resultados das bases AB e ABC foram ruins (<0,50) e nas bases A e C, assim como ocorreu para o método Cate-Nelson, os melhores resultados foram obtidos com os modelos 55-6 e 62-13.

Tabela 7. Impacto dos modelos filtrados com o filtro probabilístico multivariado para os dados discretizados com a abordagem *Discretize*.

	Modelos	r	EMA	Discretize (3 faixas de valores balanceados)						
				A	B	C	AB	ABC	I (médio)	CV
1	55-6	0,71	1,86	0,55	0,33	0,50	0,10	0,22	0,34	0,49
2	62-13	0,71	1,90	0,55	0,33	0,50	0,10	0,22	0,34	0,49
3	14	0,66	2,65	0,32	0,77	0,33	0,00	0,28	0,34	0,73
4	54-5	0,42	2,56	0,31	0,25	0,32	0,23	0,31	0,29	0,13
5	56-7	0,72	1,80	0,40	0,28	0,42	0,05	0,12	0,25	0,59
6	67-18	0,71	1,88	0,26	0,33	0,30	0,10	0,22	0,25	0,32
7	63-14	-	-	0,45	0,00	0,38	0,34	0,00	0,24	0,83
8	57-8	0,84	1,70	0,38	0,00	0,39	0,21	0,00	0,20	0,88
9	61-12	0,82	1,71	0,42	0,00	0,44	0,00	0,00	0,17	1,23
10	51-2	0,82	1,77	0,42	0,00	0,44	0,00	0,00	0,17	1,23
11	64-15	-	-	0,42	0,00	0,44	0,00	0,00	0,17	1,23
12	35	0,70	2,58	0,48	0,00	0,35	0,00	0,00	0,17	1,25
13	58-9	0,66	2,28	0,42	0,00	0,44	0,00	0,00	0,17	1,23
14	60-11	0,62	2,40	0,42	0,00	0,44	0,00	0,00	0,17	1,23
15	65-16	0,67	1,98	0,47	0,00	0,00	0,00	0,00	0,09	2,00
16	53-04	0,63	2,42	0,47	0,00	0,00	0,00	0,00	0,09	2,00
17	59-10	0,63	2,50	0,47	0,00	0,00	0,00	0,00	0,09	2,00
18	43	0,62	2,13	0,47	0,00	0,00	0,00	0,00	0,09	2,00
19	22	0,58	2,28	0,47	0,00	0,00	0,00	0,00	0,09	2,00
20	13	0,84	1,79	0,00	0,00	0,00	0,00	0,00	0,00	0,00
21	29	0,82	1,63	0,00	0,00	0,00	0,00	0,00	0,00	0,00
22	66-17	0,82	1,77	0,00	0,00	0,00	0,00	0,00	0,00	0,00
23	52-3	0,81	1,74	0,00	0,00	0,00	0,00	0,00	0,00	0,00
24	48	0,79	1,57	0,00	0,00	0,00	0,00	0,00	0,00	0,00
25	12	0,79	2,07	0,00	0,00	0,00	0,00	0,00	0,00	0,00
26	39	0,78	2,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00
27	33*	0,76	2,48	0,00	0,00	0,00	0,00	0,00	0,00	0,00
28	45*	0,76	2,48	0,00	0,00	0,00	0,00	0,00	0,00	0,00
29	44	0,74	2,46	0,00	0,00	0,00	0,00	0,00	0,00	0,00
30	31	0,74	2,46	0,00	0,00	0,00	0,00	0,00	0,00	0,00
31	40	0,73	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00
32	37	0,70	2,36	0,00	0,00	0,00	0,00	0,00	0,00	0,00
33	46	0,63	2,85	0,00	0,00	0,00	0,00	0,00	0,00	0,00
34	24	0,60	2,15	0,00	0,00	0,00	0,00	0,00	0,00	0,00
35	30	0,53	2,43	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Fonte: autoria própria.

Utilizando os filtros estatísticos, primeiramente foi aplicado o filtro baseado no teste de hipóteses sobre a correlação de uma população. Este utiliza a correlação de Pearson para verificar se há correlação linear significativa entre os valores de gesso preditos e os valores reais aplicados. O quadro com todos os resultados está no apêndice B.

Na tabela 8 constam somente os modelos que apresentaram coeficiente de correlação significativo ao nível de confiança 0,20 e um coeficiente de variação inferior a 25% para as cinco bases de dados. Os modelos aparecem ordenados pelo coeficiente de correlação médio decrescente, pelo coeficiente de variação crescente, pelo coeficiente de correlação calculado

pelo *Weka* decrescente e, por fim, pelo erro médio absoluto crescente. Observa-se que há uma tendência de que modelos que são formados por mais termos (mais variáveis explicativas) apresentem maiores valores para o coeficiente de correlação, em acordo com Ning, Kumar e Steinbach (2013). Não é possível traçar um comparativo entre os resultados da tabela 8 e a aplicabilidade além da preferência pelos modelos com maior valor de coeficiente de correlação. Porém, como já discutido, no domínio agronômico são preferíveis modelos com menos variáveis explicativas.

Tabela 8. Melhores resultados do filtro estatístico baseado em teste de hipóteses sobre a correlação de uma população.

	Modelos	r	EMA	Teste de Hipóteses Sobre a Correlação de uma População						
				A	B	C	AB	ABC	$\bar{r}$	CV
1	Modelo36	0,72	2,34	0,69	0,62	0,79	0,66	0,68	0,69	0,08
2	Modelo50-1	0,84	1,65	0,66	0,60	0,67	0,64	0,60	0,64	0,04
3	Modelo57-8	0,84	1,70	0,55	0,64	0,72	0,61	0,61	0,63	0,09
4	Modelo39	0,78	2,10	0,53	0,64	0,76	0,60	0,60	0,63	0,13
5	Modelo63-14	-	-	0,51	0,67	0,76	0,61	0,62	0,63	0,13
6	Modelo56-7	0,72	1,80	0,48	0,66	0,80	0,58	0,57	0,62	0,18
7	Modelo7	0,51	3,01	0,47	0,67	0,78	0,59	0,59	0,62	0,17
8	Modelo67-18	0,71	1,88	0,53	0,71	0,65	0,62	0,55	0,61	0,11
9	Modelo66-17	0,82	1,77	0,54	0,62	0,73	0,59	0,55	0,60	0,12
10	Modelo55-6	0,71	1,86	0,48	0,63	0,80	0,55	0,54	0,60	0,19
11	Modelo23	0,60	2,13	0,38	0,71	0,79	0,57	0,57	0,60	0,23
12	Modelo40	0,73	2,30	0,48	0,61	0,75	0,56	0,58	0,59	0,15
13	Modelo35	0,70	2,58	0,49	0,59	0,74	0,55	0,58	0,59	0,14
14	Modelo16	0,43	3,27	0,50	0,58	0,72	0,55	0,58	0,59	0,12
15	Modelo61-12	0,82	1,71	0,51	0,59	0,71	0,56	0,54	0,58	0,12
16	Modelo51-2	0,82	1,77	0,51	0,59	0,71	0,56	0,54	0,58	0,12
17	Modelo52-3	0,81	1,74	0,50	0,59	0,72	0,55	0,53	0,58	0,13
18	Modelo58-9	0,66	2,28	0,50	0,59	0,71	0,56	0,54	0,58	0,12
19	Modelo60-11	0,62	2,40	0,50	0,59	0,71	0,56	0,54	0,58	0,12
20	Modelo24	0,60	2,15	0,41	0,63	0,77	0,53	0,56	0,58	0,21
21	Modelo17	0,48	3,19	0,42	0,64	0,77	0,54	0,56	0,58	0,20
22	Modelo65-16	0,67	1,98	0,42	0,64	0,76	0,53	0,47	0,57	0,22
23	Modelo22	0,58	2,28	0,43	0,64	0,75	0,54	0,47	0,57	0,21
24	Modelo29	0,82	1,63	0,42	0,59	0,74	0,51	0,53	0,56	0,19
25	Modelo62-13	0,71	1,90	0,43	0,60	0,80	0,49	0,46	0,56	0,24
26	Modelo64-15	-	-	0,45	0,61	0,73	0,53	0,47	0,56	0,18
27	Modelo48	0,79	1,57	0,43	0,57	0,77	0,50	0,42	0,54	0,24
29	Modelo33	0,76	2,48	0,47	0,53	0,78	0,52	0,40	0,54	0,24
30	Modelo45	0,76	2,48	0,47	0,53	0,78	0,52	0,40	0,54	0,24
31	Modelo44	0,74	2,46	0,42	0,57	0,77	0,50	0,42	0,54	0,24
32	Modelo31	0,74	2,46	0,42	0,57	0,77	0,50	0,42	0,54	0,24
35	Modelo10	0,53	3,01	0,59	0,50	0,68	0,55	0,36	0,54	0,20
36	Modelo42	0,42	3,24	0,46	0,56	0,73	0,50	0,42	0,54	0,20
37	Modelo53-4	0,63	2,42	0,46	0,54	0,72	0,50	0,40	0,52	0,21
38	Modelo11	0,56	3,04	0,71	0,42	0,53	0,55	0,40	0,52	0,21
39	Modelo6	0,56	3,04	0,71	0,42	0,53	0,55	0,40	0,52	0,21
41	Modelo49	0,64	2,07	0,40	0,55	0,68	0,50	0,44	0,51	0,19

42	Modelo54-5	0,42	2,56	0,57	0,47	0,46	0,52	0,47	0,50	0,09
47	Modelo12	0,79	2,07	0,57	0,40	0,43	0,48	0,33	0,44	0,18
48	Modelo59-10	0,63	2,50	0,58	0,40	0,42	0,48	0,34	0,44	0,18
50	Modelo46	0,63	2,85	0,49	0,38	0,52	0,43	0,31	0,43	0,18
51	Modelo8	0,36	3,51	0,49	0,38	0,52	0,43	0,31	0,43	0,18

Fonte: autoria própria.

Ao aplicar isoladamente o filtro baseado no teste de hipóteses de postos de sinais de Wilcoxon foram obtidos os resultados descritos no apêndice C. Somente os modelos 13, 26, 42, 43, 55-06, 56-07, 57-08, 62-13 falharam em rejeitar a hipótese nula que era "Não há diferença entre os valores preditos e os valores reais", a um nível de significância de 0,01 em todas as bases de dados e, por este motivo, foram selecionados. A tabela 9 mostra os 6 dos 8 modelos selecionados pelo teste de Wilcoxon com os valores de coeficiente de correlação calculados pelo teste sobre a correlação. Os modelos 26 e 43 não aparecem na relação porque não apresentaram correlação linear significativa a $\alpha = 0{,}20$. Essa tabela está ordenada pelo coeficiente de correlação linear médio decrescente e pelo coeficiente de variação crescente.

A tabela 9 mostra que os melhores resultados, em média, foram para os modelos 56-07 (gesso = 0,032 x época + 0,2842 x Ca_{AB} – 0,2613 x Mg_{AB} – 0,5218 x K_{AB} + 0,1077) e 55-06, ainda que apresentem valores de erro médio elevados. O modelo 56-07 também se mostrou melhor sobre os dados das bases B e C. O modelo 55-06 apresentou resultados medianos de *r* na base de dados B, e na base C, apesar de um valor alto (0,80) o erro também foi elevado. Os demais resultados foram todos ruins, à exceção das predições dos modelos 64-15 (gesso = 0,1186 x época + 0,2663 x $CaCTCe_A$ + 0,0788 x $CaCTCe_C$ – 0,043 x $MgCTCe_D$ – 17,5739), 42 e 53-04 sobre os dados da base C que apresentaram bons valores para o coeficiente de correlação, porém, para os dois últimos modelos, valores de EMA foram elevados.

Tabela 9. Melhores regras selecionadas primeiramente com o teste sobre a correlação a $\alpha = 0{,}20$ e depois com teste de postos de sinais de Wilcoxon a $\alpha = 0{,}01$.

M	BD	A		B		C		AB		ABC		$\bar{r}$		$\overline{EMA}$	
		r	EMA	r	EMA	r	EMA	r	EMA	r	EMA	$\bar{r}$	CV	$\overline{EMA}$	CV
57-08	ABC	0,55	5,90	0,64	6,20	0,72	4,20	0,61	6,10	0,61	5,80	0,63	0,09	5,64	0,13
56-07	B	0,48	3,50	0,66	2,30	0,80	2,90	0,58	2,80	0,57	2,80	0,62	0,17	2,86	0,13
55-06	B	0,48	3,90	0,63	2,50	0,80	4,30	0,55	3,00	0,54	3,20	0,60	0,18	3,38	0,19
62-13	B	0,43	4,70	0,60	2,70	0,80	7,40	0,49	3,50	0,46	4,20	0,56	0,24	4,50	0,36
64-15	B	0,45	4,70	0,61	3,30	0,73	2,10	0,53	3,90	0,47	3,60	0,56	0,18	3,52	0,24
42	ABC	0,46	3,80	0,56	2,60	0,73	2,60	0,50	3,10	0,42	3,00	0,53	0,20	3,02	0,15
53-04	AB	0,46	3,90	0,54	2,60	0,72	2,90	0,50	3,10	0,40	3,10	0,52	0,20	3,12	0,14

Fonte: autoria própria.

Os resultados do filtro probabilístico multivariado (algoritmo RegFilter) foram combinados com os resultados dos filtros estatísticos, Os resultados são mostrados na tabela 10.

Tabela 10. Melhores regras selecionadas primeiramente com o teste sobre a correlação a $\alpha = 0{,}20$ e depois com teste de postos de sinais de Wilcoxon a $\alpha = 0{,}01$ e ordenadas pelo índice de Impacto

	Modelo	r	EMA	$\bar{r}$	$\overline{EMA}$	$\bar{I}$
1	55-6	0,71	1,86	0,6	3,38	0,56
2	62-13	0,71	1,9	0,56	4,5	0,56
3	53-04	0,63	2,42	0,52	3,12	0,56
4	56-7	0,72	1,8	0,62	2,86	0,45
5	64-15	-	-	0,56	3,52	0,45
6	57-8	0,84	1,7	0,63	5,64	0,38

Fonte: autoria própria.

A tabela 10 mostra que os modelos 55-6, 62-13 e 53-4 são aqueles que apresentam melhor resultado médio após a combinação dos filtros. Além disso, o interesse sobre eles é equivalente (0,56). Do ponto de vista do coeficiente de correlação linear, o modelo 55-6 apresenta o melhor resultado (0,6). A única ressalva está na média do erro médio absoluto, que para todos os modelos se mostra impeditivo para generalizações.

Sobre os modelos 56-07, 57-08 (gesso = 0,1168 x época + 0,0051 x $CaCTCe_{AB}$ – 0,3459 x $MgCTCe_{AB}$ – 0,3584 x $KCTCe_{AB}$ + 9,7584) e 64-15, selecionados pelos filtros estatísticos, cabe ressaltar que não foram selecionados pelo filtro probabilístico multivariado tendo em vista que o impacto médio foi menor do que 0,5. Eles contêm os atributos: gesso, época, Ca_{AB}, Mg_{AB} e K_{AB} (56-07 e 57-08) e gesso, época, $CaCTCe_A$, $CaCTCe_C$ e $MgCTCe_D$. O primeiro inclui o K e os demais o KCTCe, que segundo Silva (2013) não deveriam constar nos modelos por não apresentarem um comportamento regular em relação à lixiviação como apontado pelos trabalhos de Souza e Ritchey (1986) e Caires *et al*. (1998, 2004). Além disso, o modelo 64-15 apresenta atributos de três camadas diferentes do solo, o que dificulta sua aplicação prática. Portanto, os modelos foram corretamente eliminados pelo filtro probabilístico multivariado e incorretamente selecionados pelos filtros estatísticos baseados em testes de hipóteses.

6 CONSIDERAÇÕES FINAIS

Neste trabalho foram apresentadas duas abordagens para analisar a interessabilidade, uma estatística com testes de hipóteses sobre a correlação de Pearson e de postos de sinais de Wilcoxon e outra probabilística multivariada com a utilização de redes bayesianas. Foi desenvolvido um algoritmo chamado *RegFilter* bem como uma nova medida denominada Impacto, a qual faz uso da Sensitividade de redes bayesianas para verificar o grau de interesse de um especialista sobre um modelo de regressão. Neste trabalho o interesse foi baseado em um critério de utilidade/aplicabilidade dos modelos.

Na análise de interessabilidade, a primeira abordagem aplicada fez uso de análise subjetiva na qual o conhecimento do domínio foi representado em uma rede bayesiana e em regras sobre correlação entre atributos. Os resultados obtidos apontaram os modelos 55-06 (gesso = 0,0396 x época + 0,2891 x Ca_{AB} – 0,2809 x Mg_{AB} – 1,8504), 62-13 (gesso = 0,0485 x época + 0,25769 x Ca_A – 0,1783 x Mg_B – 5,3371), 53-04 (gesso = 0,0664 x época + 0,2748 x $CaCTCe_{AB}$ – 12,714), 59-10 (gesso = 0,051 x época + 0,2423 x $CaCTCe_C$ – 7,688), 43 (gesso = 0,0312 x época + 0,237 x $CaCTCe_A$ – 11,6847) e 22 (gesso = 0,0211 x época + 0,1917 x $CaCTCe_A$ + 0,0545 x $CaCTCe_D$ – 10,7622) como mais interessantes sob uma abordagem de utilidade. Isto utilizando o método de discretização Cate-Nelson. Com o método Discretize utilizando três faixas de valores balanceados os resultados foram insatisfatórios. Dentre estes modelos, observou-se que os modelos 55-06 e 62-13 apresentaram excelentes resultados em todas as áreas investigadas (A, B, C, AB, ABC). Houve um destaque maior para as áreas A e C, as quais apresentavam valores de acidez alta e baixa, respectivamente, e solos com textura média e argilosa. Por outro lado, na região B observou-se um melhor desempenho dos modelos 53-04, 59-10, 43 e 22, nesta região a textura do solo era argilosa e a acidez média.

O filtro probabilístico multivariado classificou somente seis modelos (I médio >= 0,5) dentre 67 (9,0%) sendo que destes somente o 22 apresenta uma dificuldade para aplicação prática pois apresenta o atributo CaCTCe nas camadas A (0-10cm) e D (40-60cm). Sobre os demais, todos poderiam ser considerados agronomicamente coerentes.

Utilizando os filtros estatísticos, chegou-se aos modelos 57-08 (gesso = 0,1168 x época + 0,0051 x $CaCTCe_{AB}$ – 0,3459 x $MgCTCe_{AB}$ – 0,3584 x $KCTCe_{AB}$ + 9,7584), 56-07 (gesso = 0,032 x época + 0,2842 x Ca_{AB} – 0,2613 x Mg_{AB} – 0,5218 x K_{AB} + 0,1077), 55-06 e 62-13 (gesso = 0,0485 x época + 0,25769 x Ca_A – 0,1783 x Mg_B – 5,3371) como mais interessantes. Neste caso a utilidade foi verificada pela aplicabilidade dos modelos sobre os

dados das regiões de estudo. Isso foi avaliado com a análise da correlação da população e uma comparação entre as distribuições de probabilidade do gesso aplicado e do gesso predito. Novamente as regras 55-06 e 62-13 apresentaram bons resultados e podem ser apontadas como as mais interessantes dentre as 67. O resultado destes filtros mostrou que regras consideradas úteis sob o conhecimento prévio (Impacto) tais como 53-04, 59-10, 43 e 22, podem ser de difícil generalização por questões relacionadas à acurácia (EMA). Além disso, dois modelos podem ser considerados incorretamente classificados, o 64-15 (gesso = 0,1186 x época + 0,2663 x $CaCTCe_A$ + 0,0788 x $CaCTCe_C$ – 0,043 x $MgCTCe_D$ – 17,5739) e o 57-8 (gesso = 0,1168 x época + 0,0051 x $CaCTCe_{AB}$ – 0,3459 x $MgCTCe_{AB}$ – 0,3584 x $KCTCe_{AB}$ + 9,7584), conforme discutido no quadro 3.

Conclui-se, portanto, que o filtro probabilístico multivariado obteve um melhor desempenho na tarefa de analisar a interessabilidade de regras de regressão e que ele se mostrou adequado para aplicação no domínio da análise de dosagem de gesso agrícola. Os modelos mais interessantes foram o 55-06 e 62-13 para as bases de dados A e C e o 53-04 para a base de dados B. Se forem considerados os resultados médios seria apontado como o modelo mais interessante o 55-06. Verificando os atributos envolvidos nos modelos observa-se que o Ca (e o CaCTCe) são os principais, seguidos do Mg (e o MgCTCe). Este resultado está de acordo com a discussão de Silva (2013) e de Guimarães (2015) onde o Ca (e o CaCTCe) também teve importância destacada.

7 TRABALHOS FUTUROS

Como sugestão de trabalhos futuros seria interessante explorar:

- Outros modelos de regressão, não somente lineares, sobre as bases de dados de necessidade de gesso agrícola e submetê-los à abordagem probabilística multivariada de análise de interessabilidade.
- Investigar a aplicabilidade das regras selecionadas neste trabalho para predição da necessidade de gesso em outras áreas com mesmas características do solo.
- Expandir o software desenvolvido para incorporar novos filtros para a regressão e também permitir a análise de modelos de classificação.

8 REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, R.; IMIELINSKI, T.; SWANI, A. Mining Association Rules Between Sets of Items in Large Databases. **Proceedings of the 1993 ACM SIGMOD Conference**, Washington DC, maio 1993.

BELLANDI, A. et al. Ontology-driven association rule extraction: a case study. **Workshop Context and Ontologys: Representation and Reasoning**, p. 1-10, 2007.

BIE, T. D. Maximum Entropy models and subjective interestingness: an application to tiles in binary databases. **Data Mining Knowledge Discovery**, p. 407-446, 2011.

BIE, T. D.; KONTONASIOS, K.-N.; SPYROPOULOU, E. A framework for mining interesting pattern sets. **SIGKDD Explorations**, Washington, v. 12, p. 92-100, 2010.

BRAGA, J. M. Importância do Modelo Matemático da Determinação do Nível Crítico de Potássio no Solo. **Pesquisa Agropecuária Brasileira**, n. 11, p. 71-75, 1976.

BRIN, S. et al. Dynamic Itemset Counting and Implication Rules for Market Basket Data, p. 255-264, 1997. Disponível em: http://dl.acm.org/citation.cfm?id=253325 acesso em 17-dez-2014.

CAIRES, E. F. et al. 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 de Ciência do Solo**, n. 22, p. 27-34, 1998.

CAIRES, E. F. et al. 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 de Ciência do Solo**, n. 23, p. 315-327, 1999.

CAIRES, E. F. et al. Lime and Gypsum Application on the Wheat Crop. **Scientia Agricola**, v. 59, n. 2, p. 357-364, abr/jun 2002.

CAIRES, E. F. et al. Alterações Químicas do Solo e Resposta da Soja ao Calcário e Gesso Aplicados na Implantação do Sistema Plantio Direto. **Revista Brasileira de Ciência do Solo**, n. 27, p. 275-286, 2003.

CAIRES, E. F. et al. Alterações químicas do solo e resposta do milho à calagem e aplicação de gesso. **Revista Brasileira de Ciência do Solo**, v. 1, n. 28, p. 125-136, 2004.

CAIRES, E. F. et al. Surface application of gypsum in low acidic oxisol under no-till cropping system. **Scientia Agricola**, Piracicaba, v. 68, n. 2, p. 209-216, mar/abr 2011.

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.

CATE, R. B.; NELSON, L. A. A simple statistical procedure for particioning soil test correlation data into two classes. **Soil Science American Procedings**, v. 35, p. 658-660, 1971.

CHAN, H. Sensitivy analysis of probabilistic graphical models. **UNIVERSITY OF CALIFORNIA**, Los Angeles, 2006.

CHANG, C.-C.; LIN, C.-J. LIBSVM -- A Library for Support Vector Machines, 2014. Disponivel em: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>. Acesso em: 6 fevereiro 2015.

CHANGE VISION. Astah Community, 2015. Disponivel em: <http://astah.net/editions/community>. Acesso em: 06 janeiro 2015.

CHURCH, K. W.; HANKS, P. Word Association Norms, Mutual Information, and Lexicography. **Computational Linguistics**, v. 16, n. 1, p. 22-29, 1990.

COELHO, F. S. et al. Valor e predição do nível crítico de índices para avaliar o estado nitrogenado da batateira. **Revista Ciência Agronômica**, n. 44, p. 115-122, 2013.

COHEN, W. W. Fast effective rule induction. **Proceedings of the Twelfth International Conference on Machine Learning**, p. 115-123, 1995.

COLEMAN, N. T.; THOMAS, G. W. The basic chemistry of soil acidity. In: PEARSON, R. W.; ADAMS, F. **Soil acidity and liming**. [S.l.]: Madison, 1967. p. 1-41.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. **Second International Conference on Knowledge Discovery and Data Mining**, p. 226-231, 1996.

FAYYAD, U. M. **Advances in Knowledge Discovery and Data Mining**. 5ª. ed. Massachusetts: MIT Press, 1996.

FRANK, E.; WITTEN, I. H. Generating accurate rule sets without global optimization. **Fifteenth International Conference on Machine Learning**, p. 144-151, 1998.

FREITAS, A. A. On rule interesting measures. **Knowledge-Based Systems**, v. 12, p. 309-315, 1999.

GARCÍA, M. B.; BUENO, R. M. Mining interestingness measures for string pattern mining. **Knowledge-Based Systems**, v. 25, p. 45-50, 2012.

GENG, L.; HAMILTON, H. J. Interestingness Measures for Data Mining: A survey. **ACM Computing Surveys**, v. 38, n. 3, 2006.

GLASS, D. Confirmation Measures of Association Rule Interestingness. **Knowleadge-Based Systems**, n. 44, p. 65-77, 2013.

GRECO, S.; SLOWINSKI, R.; SZCZECH, I. Properties of rule interestingness measures and alternative approaches to normalization of measures. **Information Sciences**, v. 216, p. 1-16, 2012.

GUHA, S.; RASTOGI, R.; SHIM, K. CURE: an efficient clustering algorithm for large databases. **ACM SIGMOD 1998**, 1998.

GUILLAUME, S.; GRISSA, D.; NGUIFO, E. M. Categorization of interestingness measures for knowledge extraction. **CoRR**, 2012.

GUIMARÃES, A. M. et al. Estimating gyspsum requirement under no-till based on machine learning technique. **Revista Ciência Agronômica**, Fortaleza, CE, v. 46, n. 2, p. 250-257, 2015.

HAN, J.; PEI, J.; YIN, Y. Mining frequent patterns without candidate generation. **Proc. of the ACM SIGMOD Int'l Conf. on Management of Data**, Dallas, Texas, USA, 2000.

HOLMES, G.; HALL, M.; FRANK, E. Generating Rule Sets from Model Trees. **Lecture Notes in Computer Science**, v. 1747, p. 1-12, 1999.

JAROSZEWICZ, S. Scalable pattern mining with Bayesian network as background knowledge. **Data Mining and Knowledge Discovery**, n. 18, p. 56-100, 2008.

JAROSZEWICZ, S.; SIMOVICI, D. A. Interestingness of frequent itemsets using Bayesian networks as background knowledge. **Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining**, Seatle, USA, 2004.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. **Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence**, San Mateo, 1995.

KUTNER, M. H. et al. **Applied Linear Statistical Models**. 4. ed. [S.l.]: McGraw-Hill, 2004. 1396 p.

LARMAN, C. **Utilizando UML e padrões:** uma introdução à análise orientada a objetos e ao desenvolvimento interativo. 3. ed. Porto Alegre: Bookman, 2007.

LARSON, R.; FARBER, B. **Estatística Aplicada**. São Paulo: Pearson Prentice Hall, 2009.

LENCA, P. et al. On selecting interestingness measures for association rules: user oriented description am multiple criteria decision aid. **Computing, Artificial Intelligence and Information Management**, v. 184, p. 610-626, 2008.

MACKAY, D. J. C. Introduction to Gaussian Processes, Dept. of Physics, Cambridge University, UK., 1998.

MACQUEEN, J. B. Some methods for classification analysis of multivariate observations. **Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability**, p. 281-297, 1967.

MALHAS, R.; AGHBARI, Z. A. Interestingness filtering engine: mining bayesian networks for interesting patterns. **Expert Systems with Applications**, v. 36, p. 5137-51-45, 2009.

MANSING, G.; BRYSON, K. M. O.; REICHGELT, H. Using ontologies to facilitate post-processing of association rules by domain experts. **Information Sciences**, v. 181, p. 419-434, 2011.

MARQUES, R. R. Aplicação superficial de calcário e gesso em manejo conservacionista de solo para cultivo de amendoim e aveia branca. **Tese de Doutorado em Agronomia - Faculdade de Ciências Agronômicas da UNESP**, Botucatu, 2008.

MASOOD, A.; SOONG, S. Measuring Interestingness - Perspectives on Anomaly Detection. **Computer Engineering and Intelligent Systems**, v. 4, n. 1, p. 29-39, 2013.

MCGARRY, K. A survey of interestingness measures for knowledge discovery. **Knowledge Engineering Review**, v. 20, n. 1, p. 39-61, 2005.

MELO, M. B. D.; SILVA, L. M. S. D. **Aspectos Técnicos dos Citros em Sergipe**. Aracaju, SE: Embrapa Tabuleiros Costeiros, 2006.

MENEGHIN, M. F. S. et al. Avaliação da Disponibilidade de Nitrogênio no Solo para o Trigo em Latossolo Vermelho do Distrito Federal. **Revista Brasileira de Ciência do Solo**, n. 32, p. 1941-1948, 2008.

NEIS, L. et al. Gesso agrícola e rendimento de grãos de soja na região do sudoeste de Goiás. **Revista Brasileira de Ciências do Solo**, v. 2, n. 34, p. 409-416, 2010.

NING, T. P.; KUMAR, V.; STEINBACH, M. **Introdução ao Data Mining - mineração de dados**. [S.l.]: Ciência Moderna, 2013.

NORSYS. https://www.norsys.com/. **Norsys Software Corp. NeticaJ**, 2015. Acesso em: 20 janeiro 2015.

NORSYS, S. C. **NeticaJ Manual:** version 4.18 and higher. Vancouver, Canadá: [s.n.], 2010.

NUERNBERG, N. J.; RECH, T. D.; BASSO, C. Usos do gesso agrícola. **Boletim técnico nº 122**, Florianópolis, 2005. Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina.

OLIVEIRA, E. L.; PAVAN, M. A. Control of soil acidity in notillage system for soybean production. **Soil and Tillage Research**, v. 1-2, n. 38, p. 47-57, 1996.

PAVAN, M. A. Toxicity of aluminum to coffe seedlings grown in nutrient solution. **Soil Science Society American Journal**, p. 993-997, 1982.

PAVAN, M. A. et al. **Manual de análise química do solo e controle de qualidade.** Londrina: Instituto Agronômico do Paraná (IAPAR), 1992.

PEARL, J. **Probabilistic Reasoning in Intelligent Systems:** Networks of Plausible Inference. [S.l.]: Morgan Kaufmann, 1988.

PEARSON, K. Mathematical contributions to the theory of evolution. III Regression, heredity and panmixia. **Philosophical Transactions of The Royal Society A**, 1986. Disponível em acesso em 18-out-2013.

PORTUGAL, M. S. **Notas Introdutórias Sobre o Princípio de Máxima Verossimilhança: Estimação e Teste de Hipóteses**. Universidade Federal do Rio Grande do Sul. Porto Alegre, p. 24. 2014. Disponível em http://www.ufrgs.br/decon/publionline/textosdidaticos/Textodid04.pdf acesso em 16-jan-2015.

QUAGGIO, J. A. et al. 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.

QUINLAN, J. R. Induction of Decision Trees. **Machine Learning**, Boston, n. 1, p. 81-106, 1986.

QUINLAN, J. R. Learning with continuous classes. **Proceedings AT92**, Singapore, p. 343-348, 1992.

QUINLAN, J. R. C4.5: Programs for machine learning. São Francisco, CA, USA: Morgan Kaufmann Publishers, 1993.

RAIJ, B. V. Gesso na agricultura. **Informações agronômicas**, v. 22, n. 1, p. 26-27, 2008.

RAIJ, B. V. et al. **Recomendações de adubação e calagem para o Estado de São Paulo**. 2. ed. Campinas: Instituto Agronômico/Fundação IAC, 1996. 285 p.

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

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

ROUSSEEUW, P. J.; LEROY, A. M. **Robust Regression and Outlier Detection**. [S.l.]: Wiley, 2005.

RPROJECT, S. The R Project for Statistical Computing, 2015. Disponivel em: <http://www.r-project.org/>. Acesso em: 18 janeiro 2015.

RUSSEL, S.; NORVIG, P. **Inteligência Artificial**. 3. ed. Rio de Janeiro: Elsevier, 2013.

SAVASERE, A.; OMIECINSKI, E.; NAVATHE, S. An efficient algorithm for mining association rules in large databases. **Proc. of the 21st Conf. on Very Large Databases (VLDB'95)**, 1995. 432-444.

SHANNON, C. E. A Mathematical Theory of Communication. **The Bell System Technical Journal**, v. 27, p. 623-656, 1948.

SHEVADE, S. K. et al. Improvements to the SMO Algorithm for SVM Regression. **IEEE Transactions on Neural Networks**, n. 5, p. 1188-1193, 2000.

SILBERSCHATZ, A.; TUZHILIN, A. What makes patterns interesting in knowledge discovery systems. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 6, p. 970-974, 1996.

SILVA, K. S. D. Uso da Mineração de Dados para Extração de Conhecimento Agronômico Envolvendo o Uso de Gesso Agrícola, Ponta Grossa, p. 95, 2013. Dissertação de Mestrado - Universidade Estadual de Ponta Grossa. Ponta Grossa, 01 de março de 2013.

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 de Ciência do Solo**, v. 2, n. 32, p. 675-688, 2008.

SOUSA, D. M. G. D.; LOBATO, E.; REIN, T. A. Uso de Gesso Agrícola nos Solos do Cerrado. **Circular Técnica da Embrapa**, Planaltina, DF, p. 18, 2005.

SOUSA, D. M. G.; LOBATO, E. Correção da acidez do solo. In: SOUSA, D. M. G.; LOBATO, E. **Cerrado:** correlação do solo e adubação. Planaltina, DF: Embrapa Cerrados, 2002. p. 81-96.

SOUZA, D. M. G.; RITCHEY, K. D. Uso do gesso no solo de cerrado. **Seminário sobre uso de fosfogesso na agricultura**, Brasília, p. 119-144, 1986.

VITTI, G. C. et al. **Uso de gesso em sistemas de produção agrícola**. Piracicaba: GAPE, 2008. 104 p.

WANG, Y. **A new approach to fitting linear models in high dimensional spaces**. Hamilton, Nova Zelândia: [s.n.], 2000.

WANG, Y.; WITTEN, I. H. Induction of Model Trees for Predicting Continuous Classes. **Working Paper Series**, Hamilton, New Zeland, p. 12, 1996.

WEKA 3: Data Mining Software in Java, 2015. Disponivel em: <http://www.cs.waikato.ac.nz/ml/weka/index.html>. Acesso em: 05 Fevereiro 2015.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining:** practical machine learning tools and techniques. 3ª edição. ed. Burlington: Elsevier, 2011.

WU, T.; CHEN, Y.; HAN, J. Re-examination of interestingness measures in pattern mining: a unified framework. **Data Mining Knowledge Discovery**, v. 21, p. 371-397, 2010.

ZAKI, M. et al. New algorithms for fast discovery of association rules. **Proc. of the ACM SIGKDD Int'1 Conf. on Knowledge Discovery and Data Mining**, Newport Beach, CA, USA, 1997. 283-286.

ZAKI, M.; PARTHASARATHY, S.; LI, W. A localized algorithm for parallel association mining. **Proceedings of the ninth annual ACM symposium on Parallel algorithms and architectures SPAA**, New York, NY, USA, p. 321-330, 1997.

ZHANG, T.; RAMAKRISHNAN, R.; LIVNY, M. BIRCH: An efficient data clustering method for very large databases. **Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data**, Montreal, Canadá, p. 103-114, 1996.

APÊNDICE A. DESCRIÇÃO DO SOFTWARE DESENVOLVIDO PARA ANÁLISE DE INTERESSABILIDADE DE MODELOS DE REGRESSÃO

Este apêndice apresenta o software desenvolvido para fazer a análise de interessabilidade de modelos de regressão. São apresentados os princípios gerais da modelagem do sistema, o método Cate-Nelson que foi utilizado para aplicar discretização de dados e, finalmente, a implementação dos filtros baseados em testes estatísticos com o software R.

A.1 Modelagem geral

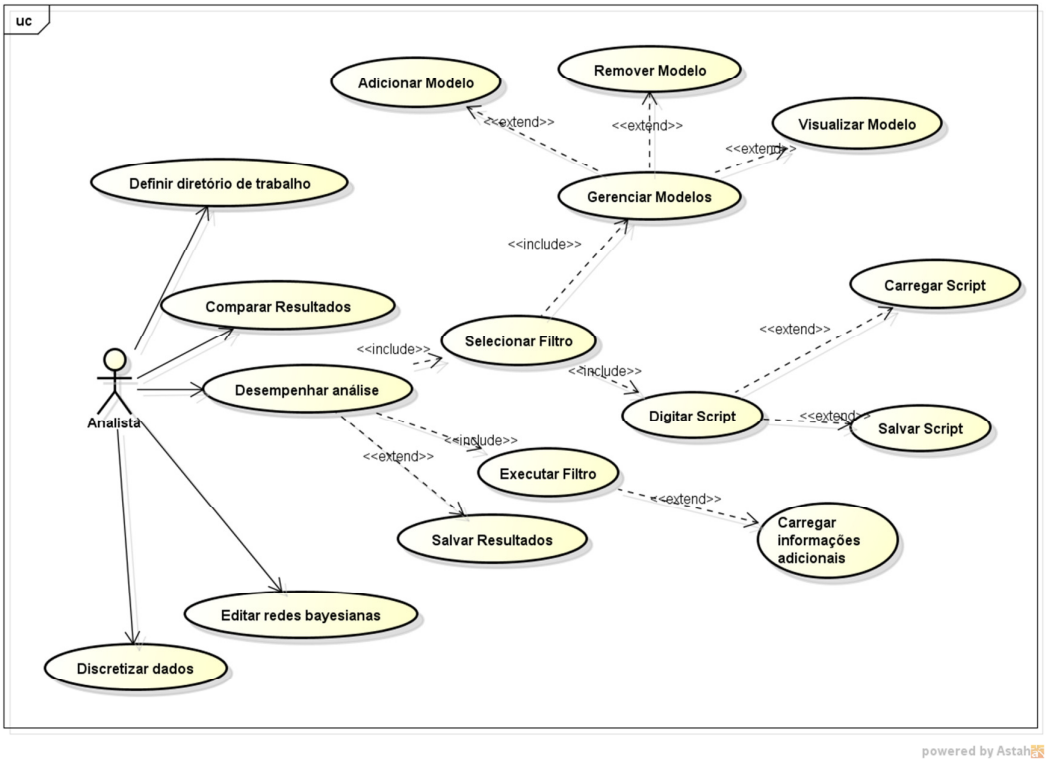
O software *Data Mining Analysis* (DMA) foi projetado para ser um ambiente de análise de interessabilidade de modelos de regressão. O desenvolvimento foi feito utilizando a linguagem Java versão 8, com o jdk versão 1.8.0_25 compilado para processadores x86 e distribuído pela Oracle. Os diagramas foram construídos utilizando o software *Astah Community* 6.9.0 (CHANGE VISION, 2015).

Foram adotados os seguintes princípios para modelagem e implementação do DMA:

1. Configuração simplificada: o software pode ser facilmente configurado para inclusão de novos filtros sem que seja necessário alterar o código pré-existente.
2. É possível ainda acrescentar algoritmos suportados com o mínimo de programação.
3. O sistema é capaz de trabalhar com modelos obtidos com diferentes algoritmos, desde que o tipo de saída seja compatível e suportado pelo filtro. Por exemplo, é possível analisar ao mesmo tempo modelos obtidos com o *M5Rules* e com o *LeastMedSq*.
4. Foram adotadas boas práticas de programação (LARMAN, 2007), como por exemplo, a adoção dos padrões de projeto MVC, do inglês *Model-View-Controller*; *Factory; TemplateMethod* e *Singleton*.
5. Atenção aos princípios da Orientação a Objetos: abstração, reuso, encapsulamento, herança, polimorfismo, entre outros.
6. Interface gráfica simples, porém, prática e intuitiva.
7. Os erros ou situações inesperadas devem ser tratados com o lançamento de exceções específicas.

A figura 8 mostra o diagrama de casos de uso descrevendo as principais funcionalidades do DMA.

Figura 8. Diagrama de casos de uso – Visão Geral



Fonte: autoria própria.

O quadro 6 apresenta uma breve descrição dos casos de uso da figura 8 destacando as funcionalidades do sistema.

Quadro 6. Descrição dos casos de uso.

Caso de Uso	Descrição
Definir diretório de trabalho	Permite que o analista defina um diretório padrão para trabalhar. Todos os arquivos utilizados pelo DMA serão procurados primeiramente neste diretório.
Desempenhar análise	Caso de uso que descreve uma nova análise. Nele deve-se obrigatoriamente selecionar um filtro dentre os disponíveis, fornecer os modelos (*buffers* de execução do Weka em arquivo texto), digitar um *script* de execução compatível com filtro escolhido e executar o filtro.
Selecionar filtro	O analista escolhe um filtro de interessabilidade de regras. Atualmente encontram-se disponíveis o *RegFilter* (seção 3.3.1) e o *StatisticalFilter* (3.3.2).
Gerenciar modelos	Deve ser utilizado para incluir novos modelos, remover modelos já adicionados ou simplesmente visualizar os modelos já carregados.

Adicionar modelo	Permite que o analista inclua novos modelos (em arquivos texto). Os modelos são interpretados e um resumo da execução do interpretador é mostrado.
Remover modelo	Utilizado para remover da memória modelos já carregados e interpretados.
Visualizar modelo	Permite visualizar em uma janela o modelo lido e interpretado.
Digitar script	Fornece uma janela na qual o analista digita um *script* e no qual informa os parâmetros de execução do filtro. Esta janela fornece uma opção com a qual é possível ver um modelo.
Carregar script	Um *script* salvo anteriormente pode ser carregado na janela de edição.
Salvar script	Utilizado para salvar um *script* de execução em disco rígido. Isso permite construir e executar diferentes experimentos de forma rápida.
Executar filtro	Coloca o filtro em execução. Só pode ser executado após um *script* ter sido digitado e interpretado.
Carregar informações adicionais	Ao iniciar a execução de um filtro os dados adicionais como bases de dados, redes bayesianas ou outros arquivos são carregados. Se isso não for possível a execução é interrompida e uma mensagem de erro apropriada é mostrada.
Salvar resultados	Permite que o resultado da execução de um filtro seja salvo no disco rígido.
Comparar resultados	O analista pode carregar arquivos com os resultados e fazer comparações. No momento são mostrados apenas os arquivos com os resultados.
Editar redes bayesianas	Neste caso de uso o analista pode construir suas redes bayesianas editando os nodos, os links, as probabilidades condicionais e executando aprendizado a partir de um arquivo de casos com o algoritmo *Expectation Maximization* (EM).
Discretizar dados	Permite utilizar discretizadores de dados. Atualmente somente o discretizador baseado no método Cate-Nelson (1971) se encontra disponível.

Fonte: autoria própria.

Os componentes do sistema encontram-se na figura 9 sendo que o sistema foi dividido em sete módulos. O módulo “interface gráfica” contém todas as janelas. Nele a entrada de dados e a saída dos resultados foram padronizadas, com isso, ao programar um filtro, por exemplo, deve-se estar atento ao padrão de entrada e saída das informações.

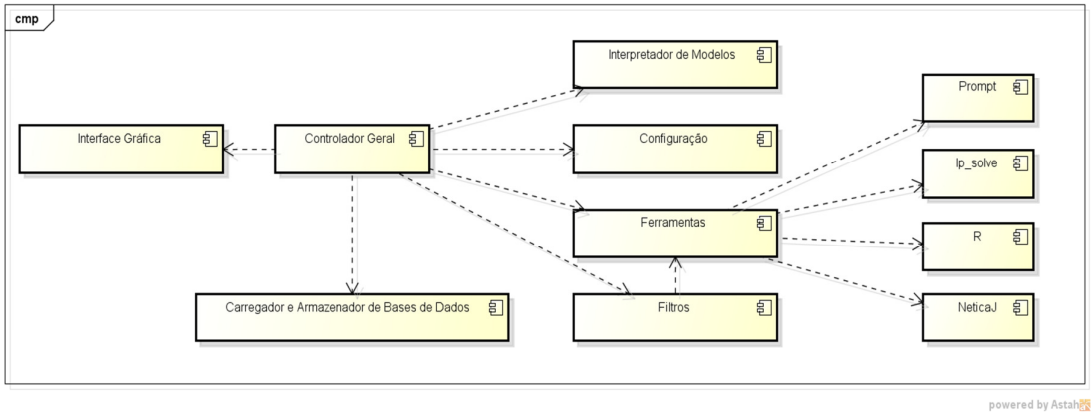
O “controlador geral” é utilizado para gerenciar o funcionamento do software integrando a interface gráfica às ferramentas, aos filtros e aos dados armazenados no disco rígido. O “carregador e armazenador de bases de dados” é um módulo que oferece as funcionalidades de ler e escrever arquivos de texto nos formatos *txt, csv, cas, arff, dne e lp* no disco rígido. O módulo “interpretador de modelos” contém as funcionalidades necessárias para ler resultados da execução de algoritmos no *Weka*. O “módulo de configuração” contém os arquivos XML com os parâmetros de funcionamento do sistema.

No módulo “ferramentas” encontram-se funcionalidades para editar redes bayesianas e discretizar dados. Outras ferramentas poderão ser incorporadas neste módulo, por exemplo, outros métodos de discretização.

Em “filtros” encontram-se os filtros de interessabilidade de regras. O sistema foi projetado para ser expansível, com isso, novos filtros podem ser adicionados neste módulo.

Por fim, é possível utilizar outros softwares para desempenhar algumas tarefas. No momento os softwares suportados são o *Prompt* (para operações sobre arquivos) e o *Rscript* para execução de scripts no *R Statistical Software* (RPROJECT, 2015), contudo, outros softwares que possam ser acessados ao nível de permissão do usuário do sistema operacional ou que estejam no path do *Windows* podem ser lançados conforme a necessidade.

Figura 9. Diagrama de componentes do DMA



Fonte: autoria própria.

O *Weka* foi escolhido porque possui um amplo conjunto de algoritmos para mineração de dados (WITTEN; FRANK; HALL, 2011), além de ser simples de utilizar e permitir exportar o resultado das aplicações dos algoritmos em arquivos de texto que podem ser lidos e interpretados pelo DMA.

O DMA suporta os algoritmos *M5Rules*, *SMOReg*, *LibSVM*, *SimpleLinearRegression*, *PaceRegression*, *LeastMedSq* e *LinearRegression* porque todos possuem uma saída que pode ou não conter restrições na forma de *IF ... THEN* e uma ou mais equações formadas por somas de termos, onde estes termos são formados por um coeficiente (positivo ou negativo) que multiplica uma variável do problema elevada a 1 ou 0.

O DMA é capaz de efetuar cálculos e comparações com os valores destas medidas, além de ser capaz de ler o modelo de regressão e efetuar cálculos para descobrir o valor da variável meta a partir dos valores das variáveis preditoras.

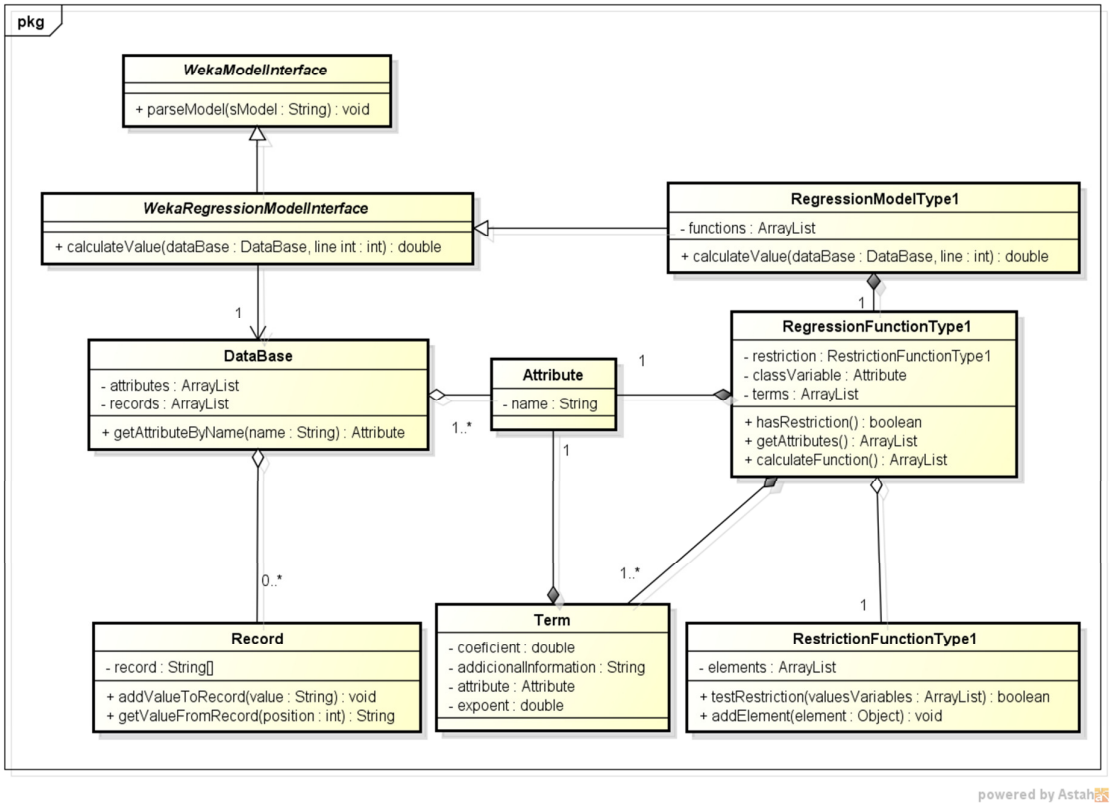
A.2 Principais diagramas de classes

Como o sistema deve ser capaz de suportar diferentes tipos de modelos, foi definida uma classe abstrata *WekaModelInterface* a qual possui o método *parseModel* e que representa a mais alta abstração para os modelos. Além disso, foi definida outra classe abstrata, que herda de *WekaModelInterface* e que representa os modelos de regressão, denominada, *WekaRegressionModelInterface.* Esta possui, além do método herdado, o método *calculateValue* que é usado para calcular o valor da variável meta dados os valores das variáveis preditoras. A figura 10 apresenta o diagrama de classes que mostra estes relacionamentos e a forma como foi modelado o *RegressionModelType1*.

Para permitir que o sistema pudesse trabalhar com vários filtros e que estes pudessem ser criados sem alterar outras partes do sistema foi definida uma classe abstrata chamada *FilterInterface*. Todos os filtros devem implementar esta interface.

Do mesmo modo, para que todos os filtros forneçam resultados padronizados, para que estes resultados possam ser exibidos em uma interface gráfica única e para que possam ser comparados automaticamente é necessário implementar a classe abstrata *ResultInterface*.

Figura 10. Diagrama de classes mostrando a representação dos modelos de regressão no DMA.

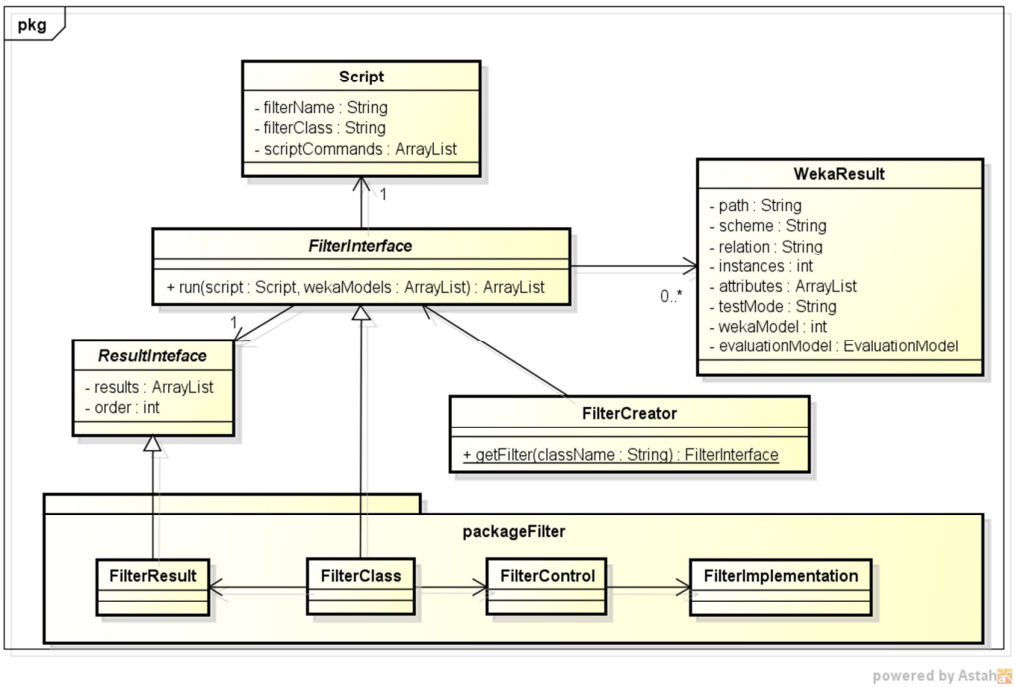


Fonte: autoria própria.

A figura 11 mostra o esquema básico para criação de novos filtros tomando por base as classes abstratas que necessitam de implementação. Nela a classe *WekaResult* modela os dados de um modelo de regressão ou classificação e também fornece um modo de acessar os valores das medidas de avaliação calculadas pelo *Weka*.

Já a classe *FilterCreator*, com seu método *getFilter( )* fornece a instância de um objeto *Filter* correspondente à string passada por parâmetro. Esta string contém o texto informado no campo <*filterclass*> do arquivo *filters.xml*.

Figura 11. Esquema básico para criação de novos filtros



Fonte: autoria própria.

A.3 Método Cate-Nelson

A discretização de dados consiste em transformar um atributo contínuo em categórico. Esta tarefa envolve duas sub-tarefas: decidir o número de categorias e determinar os pontos de corte para cada categoria (NING; KUMAR; STEINBACH, 2013).

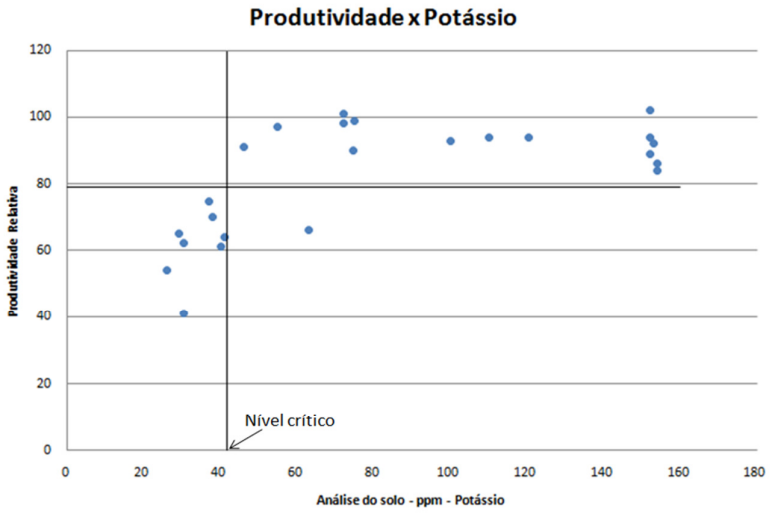
A justificativa para a adoção do método Cate-Nelson se deve ao fato de que muitos testes de solos em laboratórios com propósitos de fazer recomendações de fertilizantes necessitam dividir os resultados em duas ou mais categorias. Ele foi adotado, por exemplo, por Meneguin *et al*. (2008) em uma análise que objetivava calibrar métodos de determinação de nitrogênio em solos do cerrado para a cultura do trigo; por Braga (1976) onde o objetivo era determinar o nível crítico de potássio em 20 solos do Estado de Minas Gerais em relação

ao crescimento vegetal relativo, e por Coelho *et al*. (2013) onde o objetivo era determinar os níveis críticos de nitrogênio na batateira de modo a estimar o estado de nitrogênio na planta com base na produtividade relativa. Os passos do método Cate-Nelson são:

1. Os dados devem ser ordenados em um vetor baseado no ranqueamento dos valores de X. Os pares devem ser mantidos nesta ordem durante a análise.
2. Deve-se iniciar com o valor de X que corresponda a dois ou mais pontos no lado esquerdo de uma linha divisória dos dados. A figura 12 mostra uma linha divisória que corresponde a um ponto crítico que se deseja encontrar. Dado o valor X, deve-se então calcular a soma dos quadrados dos desvios da média corrigida para as duas populações, à direita e à esquerda do ponto X. A soma das duas somas de quadrados corrigidas é então determinada e esta soma de quadrados agrupada é subtraída da soma de quadrados dos desvios corrigida em relação à média de todas as observações de Y. A diferença entre a soma agrupada dos quadrados e a soma de quadrados total corrigida é expressa como uma porcentagem da soma total de quadrados corrigida, ou como R^2.
3. Por um processo iterativo simples pode-se obter uma série de valores de R^2 para as divisões feitas nos vários níveis de X. O valor crítico, figura 12, é aquele para o qual R^2 é máximo.

Na figura 12 o ponto crítico pode ser entendido como aquele que melhor divide os dados sob o ponto de vista da capacidade de predição.

Figura 12. Gráfico da produtividade em relação ao potássio



Fonte: Adaptado de Cate e Nelson (1971)

A.4 Teste de Hipóteses sobre a Correlação de uma População com o software R

Utilizando o software R é possível desempenhar o teste com a função *cor.test* do pacote *stats*, a qual permite testar a associação entre amostras pareadas usando o coeficiente de correlação de Pearson. No quadro 7 é apresentado um *script* contendo os comandos necessários e nos quadros 8 e 9 as saídas para os testes sobre as predições dos modelos 1 e 2 da seção 2.2.2.1.

Quadro 7. Script em R para o teste de hipóteses sobre a correlação de uma população

```
data<-read.csv("c:\\DMAFiles\\exemplo.csv")
test<-cor.test(data[,1],data[,2],alternative="two.sided",method="pearson",exact=NULL,conf.level=0.95,
continuity=TRUE)
print(test)
```

Fonte: autoria própria.

Para desempenhar o teste para o modelo 2 seria necessário alterar o texto data[,2] para data[,3].

Com o resultado da execução observa-se que a estatística de teste ficou próxima do valor calculado com a equação 7 na seção 2.2.2.1, a diferença para menos se deve ao arredondamento do coeficiente de correlação para duas casas após a vírgula.

Quadro 8. Resultado da execução do script do quadro 7.

```
        Pearson's product-moment correlation
data:  data[, 1] and data[, 2]
t = 6.0896, df = 4, p-value = 0.003677
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.6049584 0.9946876
sample estimates:
      cor
0.9500717
```

Fonte: autoria própria.

Quadro 9. Resultado da execução do script do quadro 7 para o modelo 2

```
        Pearson's product-moment correlation
data:  data[, 1] and data[, 3]
t = 2.7331, df = 4, p-value = 0.05227
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0131955  0.9780250
sample estimates:
    cor
0.807008
```

Fonte: autoria própria.

A interpretação do resultado precisa levar em conta uma comparação entre o *p-value* e o nível de significância α.

- Se o *p-value* é menor ou igual a α, deve-se rejeitar a hipótese nula. Neste caso significa que há uma correlação linear significativa entre as variáveis.
- Se o *p-value* for maior que α deve-se falhar em rejeitar a hipótese nula, ou seja, não há correlação linear significativa entre as variáveis.

Dos quadros 8 e 9 obtêm-se que $p-value_{\text{modelo1}} = 0{,}0036$ e $p-value_{\text{modelo 2}} = 0{,}0523$ e, portanto, somente o modelo 1 apresenta correlação linear significativa ao nível de confiança 5%.

A.5 Teste de Hipótese de Postos de Sinais de Wilcoxon com o software R

No software R existe a função *wilcox.test* do pacote *stats* que permite desempenhar o teste de sinais de postos e o teste de soma de postos de Wilcoxon, também conhecido como teste *Mann-Witney*. No quadro 10 é mostrado o script em R para realizar o teste de postos de sinais e no quadro 11 é mostrado resultado do teste.

Quadro 10. Script em R para desempenhar o teste de sinais de postos de Wilcoxon no R

```
data<-read.csv("c:\\DMAFiles\\exemplo2.csv")
test<-wilcox.test(data[,1],data[,2],paired=TRUE,exact=FALSE)
print(test)
```

Fonte: autoria própria.

Quadro 11. Resultado da execução do script do quadro 10 para os dados da tabela 3

```
Wilcoxon signed rank test with continuity correction
data:  data[, 1] and data[, 2]
V = 12, p-value = 0.125
alternative hypothesis: true location shift is not equal to 0
```

Fonte: autoria própria.

No quadro 11 é possível observar que V = 12 é o valor da estatística de teste e que o *p-value* = 0,125 > 0,05 (nível de confiança). Isto significa que se deve falhar em rejeitar H_0. É importante notar que a execução do script do quadro 10 independe do nível de confiança. Este altera somente a interpretação do resultado.

APÊNDICE B. RESULTADO DO TESTE DE HIPÓTESES SOBRE A CORRELAÇÃO DE UMA POPULAÇÃO

	Modelos	r	EMA	Teste de Hipóteses Sobre a Correlação de uma População						
				A	B	C	AB	ABC	$\bar{r}$	CV
1	Modelo36	0,72	2,34	0,69	0,62	0,79	0,66	0,68	0,69	0,08
2	Modelo50-1	0,84	1,65	0,66	0,60	0,67	0,64	0,60	0,64	0,04
3	Modelo57-8	0,84	1,70	0,55	0,64	0,72	0,61	0,61	0,63	0,09
4	Modelo39	0,78	2,10	0,53	0,64	0,76	0,60	0,60	0,63	0,13
5	Modelo63-14	-	-	0,51	0,67	0,76	0,61	0,62	0,63	0,13
6	Modelo56-7	0,72	1,80	0,48	0,66	0,80	0,58	0,57	0,62	0,18
7	Modelo7	0,51	3,01	0,47	0,67	0,78	0,59	0,59	0,62	0,17
8	Modelo67-18	0,71	1,88	0,53	0,71	0,65	0,62	0,55	0,61	0,11
9	Modelo66-17	0,82	1,77	0,54	0,62	0,73	0,59	0,55	0,60	0,12
10	Modelo55-6	0,71	1,86	0,48	0,63	0,80	0,55	0,54	0,60	0,19
11	Modelo23	0,60	2,13	0,38	0,71	0,79	0,57	0,57	0,60	0,23
12	Modelo40	0,73	2,30	0,48	0,61	0,75	0,56	0,58	0,59	0,15
13	Modelo35	0,70	2,58	0,49	0,59	0,74	0,55	0,58	0,59	0,14
14	Modelo16	0,43	3,27	0,50	0,58	0,72	0,55	0,58	0,59	0,12
15	Modelo61-12	0,82	1,71	0,51	0,59	0,71	0,56	0,54	0,58	0,12
16	Modelo51-2	0,82	1,77	0,51	0,59	0,71	0,56	0,54	0,58	0,12
17	Modelo52-3	0,81	1,74	0,50	0,59	0,72	0,55	0,53	0,58	0,13
18	Modelo58-9	0,66	2,28	0,50	0,59	0,71	0,56	0,54	0,58	0,12
19	Modelo60-11	0,62	2,40	0,50	0,59	0,71	0,56	0,54	0,58	0,12
20	Modelo24	0,60	2,15	0,41	0,63	0,77	0,53	0,56	0,58	0,21
21	Modelo17	0,48	3,19	0,42	0,64	0,77	0,54	0,56	0,58	0,20
22	Modelo65-16	0,67	1,98	0,42	0,64	0,76	0,53	0,47	0,57	0,22
23	Modelo22	0,58	2,28	0,43	0,64	0,75	0,54	0,47	0,57	0,21
24	Modelo29	0,82	1,63	0,42	0,59	0,74	0,51	0,53	0,56	0,19
25	Modelo62-13	0,71	1,90	0,43	0,60	0,80	0,49	0,46	0,56	0,24
26	Modelo64-15	-	-	0,45	0,61	0,73	0,53	0,47	0,56	0,18
27	Modelo48	0,79	1,57	0,43	0,57	0,77	0,50	0,42	0,54	0,24
28	Modelo32º	0,79	2,36	0,44	0,56	0,84	0,47	0,41	0,54	0,29
29	Modelo33	0,76	2,48	0,47	0,53	0,78	0,52	0,40	0,54	0,24
30	Modelo45	0,76	2,48	0,47	0,53	0,78	0,52	0,40	0,54	0,24
31	Modelo44	0,74	2,46	0,42	0,57	0,77	0,50	0,42	0,54	0,24
32	Modelo31	0,74	2,46	0,42	0,57	0,77	0,50	0,42	0,54	0,24
33	Modelo37º	0,70	2,36	0,37	0,60	0,79	0,48	0,44	0,54	0,27
34	Modelo43º	0,62	2,13	0,37	0,61	0,79	0,48	0,45	0,54	0,27
35	Modelo10	0,53	3,01	0,59	0,50	0,68	0,55	0,36	0,54	0,20
36	Modelo42	0,42	3,24	0,46	0,56	0,73	0,50	0,42	0,54	0,20
37	Modelo53-4	0,63	2,42	0,46	0,54	0,72	0,50	0,40	0,52	0,21
38	Modelo11	0,56	3,04	0,71	0,42	0,53	0,55	0,40	0,52	0,21
39	Modelo6	0,56	3,04	0,71	0,42	0,53	0,55	0,40	0,52	0,21
40	Modelo27º	0,78	1,63	0,36	0,55	0,79	0,45	0,41	0,51	0,30
41	Modelo49	0,64	2,07	0,40	0,55	0,68	0,50	0,44	0,51	0,19
42	Modelo54-5	0,42	2,56	0,57	0,47	0,46	0,52	0,47	0,50	0,09
43	Modelo20º	0,57	3,04	0,60	0,52	0,26	0,57	0,41	0,47	0,26
44	Modelo15º	0,55	3,01	0,61	0,52	0,27	0,57	0,41	0,47	0,26
45	Modelo5º	0,55	3,01	0,61	0,52	0,27	0,57	0,41	0,47	0,26
46	Modelo28º	0,74	1,71	0,20	0,51	0,76	0,38	0,43	0,46	0,40
47	Modelo12	0,79	2,07	0,57	0,40	0,43	0,48	0,33	0,44	0,18
48	Modelo59-10	0,63	2,50	0,58	0,40	0,42	0,48	0,34	0,44	0,18
49	Modelo13º	0,84	1,79	0,57	0,53	0,08 ^{ns}	0,56	0,41	0,43	0,42

50	Modelo46	0,63	2,85	0,49	0,38	0,52	0,43	0,31	0,43	0,18
51	Modelo8	0,36	3,51	0,49	0,38	0,52	0,43	0,31	0,43	0,18
52	Modelo47º	0,70	2,54	0,68	0,27	0,24	0,49	0,39	0,41	0,38
53	Modelo9º	0,50	3,20	0,59	0,33	0,43	0,43	0,24	0,40	0,29
54	Modelo19º	0,88	1,69	0,66	0,34	0,05 ^{ns}	0,48	0,37	0,38	0,52
55	Modelo14º	0,66	2,65	0,61	0,43	0,09 ^{ns}	0,51	0,27	0,38	0,48
56	Modelo41º	0,52	3,12	0,31	0,34	0,45	0,36	0,25	0,34	0,19
57	Modelo34º	0,60	3,08	0,29	0,21	0,58	0,30	0,26	0,33	0,40
58	Modelo18º	0,30	3,81	0,48	0,12 ^{ns}	0,51	0,28	0,22	0,32	0,46
59	Modelo4º	0,36	3,65	0,51	0,30	0,24	0,35	0,17	0,31	0,37
60	Modelo2º	0,31	3,66	0,59	0,18	0,26	0,29	0,21	0,31	0,48
61	Modelo38º	0,26	3,48	0,51	0,30	0,24	0,35	0,17	0,31	0,37
62	Modelo1º	0,25	3,85	0,59	0,22	0,29	0,33	0,11	0,31	0,51
63	Modelo30º	0,53	2,43	0,12^{ns}	0,39	0,43	0,33	0,23	0,30	0,37
64	Modelo25º	0,30	2,76	0,12^{ns}	0,39	0,43	0,33	0,23	0,30	0,37
65	Modelo21º	0,20	2,84	0,19	0,37	0,24	0,27	0,20	0,25	0,26
66	Modelo3º	0,46	3,30	0,52	0,05 ^{ns}	0,23	0,17	0,02 ^{ns}	0,20	0,90
67	Modelo26º	0,35	2,76	0,02^{ns}	0,26	0,29	0,18	0,12	0,18	0,56

º Modelos removidos porque o CV é maior do que 25%.

^{ns} Não significativo a $\alpha = 0{,}20$.

APÊNDICE C. RESULTADO DA APLICAÇÃO DO FILTRO BASEADO EM TESTE DE HIPÓTESES SOBRE OS POSTOS DE SINAIS DE WILCOXON

Relação de Modelos Filtrados com o teste de Postos de Sinais de Wilcoxon																
Nível de significância		0,90					0,95					0,99				
M	BD	A	B	C	AB	ABC	A	B	C	AB	ABC	A	B	C	AB	ABC
1	A		X	X	X	X		X	X	X	X		X	X	X	X
2	A		X	X	X	X		X	X	X	X		X	X	X	X
3	A	X	X	X		X	X	X	X		X	X	X	X		X
4	A		X	X	X	X		X	X	X	X		X	X	X	X
5	A	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
6	A			X		X			X		X			X		X
7	A	X	X	X	X	X	X	X	X	X	X	X	X		X	X
8	A		X	X	X	X		X	X	X	X		X	X	X	X
9	A	X		X	X		X		X	X		X		X		
10	A	X	X	X	X	X	X	X	X	X		X	X	X	X	
11	A			X		X			X		X			X		X
12	A	X		X	X		X		X	X				X		
13	A	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
14	A	X	X	X	X		X		X	X		X		X	X	
15	A	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
16	A	X	X		X	X	X	X		X	X	X	X		X	X
17	A	X	X		X	X	X	X		X	X	X	X		X	X
18	A	X	X	X	X	X	X	X	X	X	X	X	X		X	X
19	A	X		X	X		X		X					X		
20	A		X	X	X			X	X	X				X	X	
21	B	X		X			X		X							
22	B	X	X	X	X	X	X	X	X	X		X	X	X	X	
23	B			X		X			X					X		
24	B	X		X					X					X		
25	B			X					X					X		
26	B	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
27	B	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
28	B		X		X	X		X		X	X		X		X	X
29	B	X		X	X					X					X	
30	B	X	X	X	X	X	X	X	X	X	X			X	X	X
31	C	X	X	X	X	X	X	X	X	X	X	X	X		X	X
32	C	X	X		X	X	X	X		X	X	X	X		X	X
33	C	X	X	X	X	X	X	X	X	X	X	X	X		X	X
34	C	X	X	X	X	X	X	X	X	X	X	X	X		X	X
35	C	X	X	X	X	X	X	X	X	X	X		X	X	X	X
36	ABC		X	X												
37	ABC	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
38	ABC	X		X			X		X					X		
39	ABC		X	X				X	X				X	X		

40	ABC			X					X							
41	ABC		X	X	X			X	X	X			X	X		
42	ABC	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
43	ABC	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
44	ABC	X	X	X	X	X	X	X	X	X	X	X	X		X	X
45	ABC	X	X	X	X	X	X	X	X	X	X	X	X		X	X
46	ABC		X	X		X			X		X			X		X
47	ABC		X	X	X	X		X	X	X			X	X	X	
48	ABC	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
49	ABC			X					X					X		
50-01	ABC	X	X		X	X	X	X		X	X	X	X		X	X
51-02	ABC	X	X		X	X	X	X		X	X	X	X		X	X
52-03	ABC	X	X	X	X	X	X	X	X	X	X		X	X	X	
53-04	AB	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
54-05	B	X		X	X	X	X		X		X			X		X
55-06	B	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
56-07	B	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
57-08	ABC	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
58-09	AB	X	X	X	X	X	X	X	X	X	X	X	X		X	X
59-10	AB	X	X	X	X		X	X	X	X		X		X	X	
60-11	ABC	X	X	X	X	X	X	X	X	X	X	X	X		X	X
61-12	ABC	X	X	X	X	X	X	X	X	X	X	X	X		X	X
62-13	B	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
63-14	B	X	X		X	X	X	X		X	X		X		X	X
64-15	B	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
65-16	B	X	X		X	X	X	X		X		X	X		X	
66-17	ABC	X	X		X		X	X		X			X		X	
67-18	B			X		X			X		X			X		

M	Modelo
BD	Base de dados utilizada para criação do modelo com o M5Rules
X	Indica que o modelo foi selecionado porque não há diferença entre as distribuições de probabilidade das populações ao nível de significância indicado.
	Modelos eliminados

ANEXO A. MODELOS DE REGRESSÃO OBTIDOS POR SILVA (2013)

Modelo	Equação(ões) do Modelo	R	EMA	Área
1	gesso = 0,187 * m_A – 0,264 * m_C – 0,2553 * m_D + 13,9773	0,25	3,85	A
2	IF m_C > 18 THEN gesso = 0,0167 * epoca – 0,0686 * m_C – 0,2924 * m_D + 11,0116 gesso = + 9,1429	0,31	3,66	A
3	gesso = 0,0605 * epoca – 0,3843 * m_C + 11,853	0,46	3,30	A
4	gesso = -0,3735 * m_D + 11,9611	0,36	3,65	A
5	gesso = 0,0679 * epoca + 0,1485 * Ca_CTCe_C + 0,1406 * Ca_CTCe_D – 10,2964	0,55	3,01	A
6	gesso = 0,2962 * m_A + 0,1825 * m_B – 0,2463 * m_C + 0,1843 * Ca_CTCe_B + 0,1465 * Ca_CTCe_C – 11,1576	0,56	3,04	A
7	gesso = 0,1252 * epoca + 0,3648 * m_A + 0,3968 * Ca_CTCe_A – 25,7968	0,51	3,01	A
8	gesso = 0,1721 * Ca_CTCe_B – 2,6133	0,36	3,51	A
9	gesso = 0,0682 * epoca – 0,1551 * m_C + 0,1867 * Ca_CTCe_C – 1,933	0,50	3,20	A
10	gesso = 0,0621 * epoca – 0,1621 * m_D + 0,1929 * Ca_CTCe_D – 3,0755	0,53	3,01	A
11	gesso = 0,2962 * m_A + 0,1825 * m_B – 0,2463 * m_C + 0,1843 * Ca_CTCe_B + 0,1465 * Ca_CTCe_C – 11,1576	0,56	3,04	A
12	gesso = 0,298 * Ca_CTCe_C – 9,5716	0,79	2,07	A
13	gesso = 0,3468 * Ca_CTCe_D – 13,4394	0,84	1,79	A
14	gesso = 0,0774 * epoca – 0,2197 * m_C + 0,1819 * Ca_CTCe_D – 1,1435	0.663	2,65	A
15	gesso = 0,0679 * epoca + 0,1485 * Ca_CTCe_C + 0,1406 * Ca_CTCe_D – 10,2964	0,55	3,01	A
16	gesso = 0,0638 * epoca – 0,2176 * Mg_CTCe_A – 0,169 * Mg_CTCe_B + 13,5279	0,43	3,27	A
17	gesso = 0,081 * epoca – 0,3307 * Mg_CTCe_A + 11,3686	0,48	3,19	A
18	gesso = 0,0657 * epoca – 1,7082 * K_CTCe_A + 1,999 * K_CTCe_B – 1,5917 * K_CTCe_C + 10,1414	0,30	3,81	A

19	IF Ca_CTCe_D > 50,4 THEN
	gesso = 0,6718 * m_A – 0,1817 * m_B + 0,0844 * Ca_CTCe_A + 0,057 * Ca_CTCe_C + 0,0492 * Ca_CTCe_D – 3,2753
	IF Ca_CTCe_C <= 42,95 THEN
	gesso = 0,1547 * Ca_CTCe_C – 5,1893
	gesso = 0,4101 * m_B – 1,4129
		0,88	1,69	A

20	gesso = 0,1447 * Ca_CTCe_C + 0,1547 * Ca_CTCe_D – 9,6337	0,57	3,04	A
21	gesso = -0,3401 * m_A + 0,1774 * m_B – 0,1405 * m_D + 5,8894	0,20	2,84	B
22	gesso = 0,0211 * epoca + 0,1917 * Ca_CTCe_A + 0,0545 * Ca_CTCe_D – 10,7622	0,58	2,28	B
23	gesso = 0,1881 * m_B – 0,0776 * m_D + 0,2422 * Ca_CTCe_A + 0,0652 * Ca_CTCe_B – 0,0508 * Ca_CTCe_C - 12.8042	0,60	2,13	B
24	gesso = -0,2709 * Mg_CTCe_A + 11,4538	0,60	2,15	B
25	gesso = -0,63 * K_CTCe_B + 8,5353	0,30	2,76	B

26	IF m_D <= 9.45 THEN
	gesso = -0,1056 * epoca + 0,1284 * m_B – 0,1776 * m_D + 9,7903
	IF m_C <= 13,87 AND m_C <= 11,595 AND epoca > 14 THEN
	gesso = -0,1093 * epoca + 0,2947 * m_A + 0,1094 * m_C – 0,0607 * m_D + 4,8483
	IF m_C <= 13,87 THEN
	gesso = 0,3011 * m_A – 0,2853 * m_C + 8,1864
	IF m_A <= 3,925 THEN
	gesso = 0,2889 * m_A – 0,1247
	gesso = + 3
		0,35	2,76	B

Modelo	Equação(ões) do Modelo	R	EMA	Área
27	IF Ca_CTCe_A <= 65,55 THEN gesso = 0,0272 * epoca + 0,113 * Ca_CTCe_A + 0,1187 * Ca_CTCe_B – 0,0306 * Ca_CTCe_D – 9,6271 gesso = 0,2025 * Ca_CTCe_A – 7,746	0,78	1,63	B
28	IF Ca_CTCe_A <= 65,55 AND Ca_CTCe_B <= 50,6 THEN gesso = 0,7037 * m_A – 0,034 * m_C – 0,0336 * m_D + 0,1424 * Ca_CTCe_A + 0,1101 * Ca_CTCe_B – 0,0475 * Ca_CTCe_C – 11,1357 gesso = 0,2277 * Ca_CTCe_A – 9,7936	0,74	1,71	B
29	gesso = -0,2935 * Mg_CTCe_A – 0,1156 * Mg_CTCe_B + 0,1366 * Mg_CTCe_D + 10,2784	0,82	1,63	B
30	gesso = -1,1172 * K_CTCe_B + 12,5751	0,53	2,43	B
31	gesso = 0,3108 * Ca_CTCe_A + 0,1181 * Ca_CTCe_B – 25,4056	0,74	2,46	C
32	gesso = -2,901 * pH_A + 1,7595 * Ca_A + 1,3889 * Ca_D + 12,381 * K_D + 0,1122 * Ca_CTCe_B – 3,3318	0,79	2,36	C
33	gesso = 0,2645 * Ca_CTCe_A + 0,1104 * Ca_CTCe_B – 1,1456 * K_CTCe_A – 17,5001	0,76	2,48	C
34	IF K_CTCe_C <= 4,15 THEN gesso = -2,5294 * K_CTCe_A – 0,6626 * K_CTCe_C + 17,6311 gesso = + 1,3333	0,60	3,08	C
35	gesso = -0,0748 * epoca – 0,2782 * Mg_CTCe_A – 0,191 * Mg_CTCe_B + 18,8598	0,70	2,58	C
36	gesso = -3,5949 * pH_A + 1,9049 * pH_B + 1,3388 * Al_A – 0,1233 * Ca_A + 0,1663 * Ca_D + 1,7835 * K_A + 0,2748 * Ca_CTCe_A + 0,0974 * Ca_CTCe_B + 0,1152 * Ca_CTCe_C + 0,1158 * Mg_CTCe_C – 1,0106 * K_CTCe_A – 19,1389	0,72	2,34	ABC
37	gesso = 0,2898 * Ca_CTCe_A – 14,909	0,70	2,36	ABC
38	gesso = -0,1411 * m_D + 6,9306	0,26	3,48	ABC
39	gesso = 0,7007 * m_A + 0,2733 * Ca_CTCe_A + 0,1297 * Ca_CTCe_B – 22,969	0,78	2,10	ABC
40	gesso = -0,279 * Mg_CTCe_A – 0,1605 * Mg_CTCe_B + 0,0604 * Mg_CTCe_D + 14,1788	0,73	2,30	ABC
41	IF K_CTCe_C > 3,65 AND K_CTCe_B > 4,75 THEN gesso = 0,0532 * epoca – 0,6177 * K_CTCe_A – 0,2194 * K_CTCe_B – 0,1299 * K_CTCe_C + 10,1743 gesso = -0,7559 * K_CTCe_A – 2,6851 * K_CTCe_C + 17,7733	0,52	3,12	ABC
42	gesso = 0,1036 * epoca + 0,1406 * Ca_CTCe_A + 0,1054 * Ca_CTCe_B – 11,4338	0,42	3,24	ABC
43	gesso = 0,0312 * epoca + 0,237 * Ca_CTCe_A – 11,6847	0,62	2,13	ABC
44	gesso = 0,3108 * Ca_CTCe_A + 0,1181 * Ca_CTCe_B – 25,4056	0,74	2,46	ABC
45	gesso = 0,2645 * Ca_CTCe_A + 0,1104 * Ca_CTCe_B – 1,1456 * K_CTCe_A – 17,5001	0,76	2,48	ABC
46	gesso = 0,2409 * Ca_CTCe_B – 6,7551	0,63	2,85	ABC

47	IF Ca_CTCe_D > 50,4 THEN
	gesso = 0,1095 * Ca_CTCe_D – 1,3997 * K_CTCe_D + 6,5524
	IF Ca_CTCe_C <= 42,95 THEN
	gesso = 0,1547 * Ca_CTCe_C – 5,1893
	gesso = 0,4101 * m_B – 1,4129
		0,70	2,54	ABC

48	gesso = 0,2147 * Ca_CTCe_A + 0,0936 * Ca_CTCe_B – 14,5885	0,79	1,57	ABC

Modelo	Equação(ões) do Modelo	R	EMA	Área

49	IF Ca_CTCe_A <= 65,55 AND Mg_CTCe_B <= 37,85 AND K_CTCe_B <= 9,9 AND Mg_CTCe_B <= 29,45 THEN
	gesso = 0,0718 * Ca_CTCe_A – 0,1536 * Mg_CTCe_B + 0,1154 * Mg_CTCe_D – 0,472 * K_CTCe_B + 3,2714
	IF Ca_CTCe_A <= 64,85 AND Mg_CTCe_B <= 37,85 AND K_CTCe_B <= 9,5 THEN
	gesso = 0,0675 * m_C + 0,123 * Ca_CTCe_A – 0,0999 * Mg_CTCe_B + 0,0582 * Mg_CTCe_D – 0,4392 * K_CTCe_B – 1,0634
	IF Ca_CTCe_A <= 64,85 THEN
	gesso = 0,2357 * Ca_CTCe_A + 0,067 * Mg_CTCe_D – 15,1803
	IF Mg_CTCe_D<=40,3 AND m_C <= 11,595 AND Ca_CTCe_A <= 77,9 THEN
	gesso = -0,0323 * epoca + 0,2701 * Ca_CTCe_A + 0,0598 * Mg_CTCe_D – 14,5449
	IF m_C <= 11,63 THEN
	gesso = -0,2075 * m_C + 10,4249
	gesso = + 6,5
		0,64	2,07	ABC

50-1	gesso = 0,0959 * epoca + 1,0021 * Al_AB – 0,2187 * Ca_C + 0,1141 * Mg_AB + 0,2193 * Mg_C + 1,7824 * K_C + 0,3183 * Ca_CTCe_AB + 0,2188 * Ca_CTCe_C – 0,5856 * K_CTCe_C – 29,3984	0,84	1,65	ABC
51-2	gesso = 0,0888 * epoca + 0,1482 * Ca_CTCe_AB – 0,2998 * Mg_CTCe_AB + 2,3871	0,82	1,77	ABC
52-3	gesso = 0,1675 * Ca_CTCe_AB – 0,2647 * Mg_CTCe_AB + 1,4675	0,84	1,74	ABC
53-4	gesso = 0,0664 * epoca + 0,2748 * Ca_CTCe_AB – 12,714	0,63	2,42	AB
54-5	gesso = 0,179 * Ca_C + 0,2375 * Mg_C – 0,1599 * Mg_CTCe_C + 4,257	0,42	2,56	B
55-6	gesso = 0,0396 * epoca + 0,2891 * Ca_AB – 0,2809 * Mg_AB – 1,8504	0,71	1,86	B
56-7	gesso = 0,032 * epoca + 0,2842 * Ca_AB – 0,2613 * Mg_AB – 0,5218 * K_AB + 0,1077	0,72	1,80	B
57-8	gesso = 0,1168 * epoca + 0,0051 * Ca_CTCe_AB – 0,3459 * Mg_CTCe_AB – 0,3584 * K_CTCe_AB + 9,7584	0,84	1,70	ABC
58-9	gesso = 0,0545 * epoca + 0,137 * Ca_CTCe_AB – 0,2296 * Mg_CTCe_AB + 1,6416	0,66	2,28	AB
59-10	gesso = 0,051 * epoca + 0,2423 * Ca_CTCe_C – 7,688	0,63	2,50	AB
60-11	gesso = 0,0677 * epoca + 0,1575 * Ca_CTCe_AB – 0,2815 * Mg_CTCe_AB + 1,5963	0,62	2,40	ABC
61-12	gesso = 0,0727 * epoca + 0,1494 * Ca_CTCe_AB – 0,2925 * Mg_CTCe_AB + 2,3616	0,82	1,71	ABC
62-13	gesso = 0,0485 * epoca + 0,2579 * Ca_A – 0,1783 * Mg_B – 5,3371	0,71	1,90	B
63-14	gesso = 0,108 * epoca – 0,2967 * Mg_CTCe_A – 0,1204 * Mg_CTCe_B – 0,3339 * K_CTCe_C + 15,7802	Nd	nd	B
64-15	gesso = 0,1186 * epoca + 0,2663 * Ca_CTCe_A + 0,0788 * Ca_CTCe_C – 0,043 * Mg_CTCe_D – 17,5739	Nd	nd	B
65-16	gesso = 0,0202 * epoca + 0,2174 * Ca_CTCe_A + 0,0503 * Ca_CTCe_D – 12,1985	0,67	1,98	B
66-17	gesso = 0,119 * Ca_CTCe_C – 0,2687 * Mg_CTCe_A – 0,1288 * K_CTCe_D + 5,8379	0,82	1,77	ABC
67-18	gesso = 0,121 * Ca_A + 0,3446 * Ca_D – 0,121 * Mg_C – 2,6925	0,71	1,88	B

r	Coeficiente de correlação calculado para o fragmento da base de dados utilizada na criação do modelo
EMA	Erro médio absoluto
A	Fazenda Estância dos Pinheiros
B	Fazenda Regina
C	Guarapuava
AB	Fazenda Estância dos Pinheiros + Fazenda Regina
ABC	Fazenda Estância dos Pinheiros + Fazenda Regina + Guarapuava
-XX	Refere-se ao número que a regra possui em Silva(2013). Por exemplo, o modelo 67-18 corresponde ao 18º modelo analisado naquele trabalho e é referenciado neste trabalho com o número 67.