ISSN: 2317-0840

Estimação de estrutura em Redes Bayesianas via método scoring and restrict: uma aplicação na área de agricultura

Camila S. Ozelame[1,2], Ramon Lopes[3], Anderson Ara[4]†, Francisco Louzada[1]

[1] *Universidade de São Paulo, USP.*
[2] *Universidade Federal de São Carlos, UFSCar.*
[3] *Universidade Federal do Recôncavo da Bahia, UFRB.*
[4] *Universidade Federal da Bahia, UFBA.*

Resumo: *A necessidade de uma análise preditiva assertiva aliada à interpretação das causas de seu resultado é um dos desafios encontrados nas metodologias de aprendizado de máquina. No mesmo sentido, a metodologia de Redes Bayesianas se propõe a ser uma ferramenta de investigação causal, sendo amplamente utilizada não somente para essa mas como para outras tarefas. Nesse contexto, este artigo busca, por meio da combinação de algoritmos de estimação de estrutura de Redes Bayesianas, encontrar uma arquitetura preditiva e interpretável para proporção de falhas em talhões de plantio de cana-de-açúcar. Para tanto, propõe-se um procedimento híbrido denominado scoring and restrict e verifica-se sua adequabilidade para análise de dados simulados e reais. Da análise dos dados conclui-se que houve indícios de que esta estratégia reflete em uma melhora das medidas de adequação e das possibilidades de interpretação.*

Palavras-chave: Redes Bayesianas; Estimação de estrutura; Cana-de-açúcar.

Abstract: *The need for an assertive predictive analysis combined with the causal interpretation of its result is one of the challenges found in machine learning methodologies. Bayesian Networks methodology proposes to be a tool for causal investigation and is widely used not only for this way but also for other tasks. In this context, this paper aims to find a predictive and interpretable architecture for the proportion of failures in sugarcane fields using a combined Bayesian Network structure estimation algorithm. Therefore, a hybrid procedure namely scoring and restrict is proposed as well as its suitability for analysis in artificial and real data is verified. From these analysis, we may conclude that there are an improvement in the adequacy measures and the possibilities of the model interpretation.*

Keywords: Bayesian networks; Structure learning; sugarcane.

Introdução

Observa-se uma tendência recente de produtores rurais e as indústrias de transformação dar maior atenção a aumentar a qualidade e mitigar riscos e perdas, monitorando todo o processo agronômico, desde a plantação até a colheita e distribuição do produto. Nesse cenário, o gerenciamento de risco na produção agronômica é motivo de atenção, uma vez que existem inúmeros fatores que podem estar associados ao sucesso ou fracasso da manutenção da boa saúde produtiva e financeira de uma fazenda (HARWOOD, 1999). O Brasil, um dos celeiros mundiais, tem como sua principal atividade econômica a agropecuária com diversos cultivos em todo seu território, tais como a soja, o milho, o algodão, a cana-de-açúcar, e outras culturas valorosas tanto para a exportação quanto para o consumo interno.

†Autor correspondente: anderson.ara@ufba.br.

Dentre os produtos mencionados, a cana-de-açúcar destaca-se como um dos mais versáteis, pois, além de todos seus subprodutos, é uma alternativa direta na indústria de biocombustível. No entanto, a cana-de-açúcar vem sofrendo retração na sua área de plantio desde a safra 2017/18, por diversas razões apontadas nos levantamentos quadrimestrais da Companhia Nacional de Abastecimento (CONAB, 2018). Em 2020 a CONAB aponta que, apesar da diminuição da área plantada, a produção de cana-de-açúcar aumentou em média 2,5% em relação ao exercício anterior (CONAB, 2020). Com isso, a necessidade de mitigação dos riscos do plantio dessa variedade se torna mais pertinente para seus produtores, podendo influenciar também na qualidade e no valor comercial de seus principais derivados: o etanol e o açúcar (CONAB, 2018).

Um dos aspectos desafiadores do manejo de riscos na agricultura industrial é a mitigação de falhas em lotes produtivos, foco deste artigo, mais especificamente para o plantio da cana-de-açúcar. As falhas na plantação podem ser definidas como um intervalo linear com o crescimento de nenhuma planta (STOLF, 1986); assim a definição da dimensão de falha pode variar, e, neste artigo, será considerada uma área com 30 centímetros ou mais, na maior dimensão. As causas dessas falhas podem advir de diversos aspectos e sua identificação pode auxiliar no aumento da produção, medir a qualidade do processo operacional e aumentar a rentabilidade das lavouras (ALVES et al., 2015).

Em diversos estudos, Redes Bayesianas (PEARL, 1988) são utilizadas para modelar fenômenos complexos e probabilísticos em aplicações ambientais (RASMUSSEN; MADSEN; LUND, 2013; YET et al., 2016; XUE et al., 2017; DRURY et al., 2017). Rasmussen, Madsen e Lund (2013) utilizam Redes Bayesianas (RB) como ferramenta para o gerenciamento de riscos em unidades de produção agrícola, reiterando a ideia de minimizar incertezas com intuito de manutenção de preços coerentes e condições produtivas favoráveis. Yet et al. (2016) empregam RB para analisar o desenvolvimento agronômico nas frentes de custo de projeto, benefícios e análise de risco. Xue et al. (2017) por sua vez, aplicam redes híbridas para analisar a o fluxo ambiental e a irrigação na agricultura. Drury et al. (2017) fazem um levantamento de aplicações das RB na agricultura em diversos contextos como evolução e transmissão de doenças, controle de pragas, efeitos do aquecimento global e poluição, estratégias e viabilidade da produção agronômica. Neste artigo, investigamos o uso de RB para determinar variáveis que influenciem direta ou indiretamente o percentual de falha na plantação de cana-de-açúcar com base em dados reais.

As RB, também denominadas de redes de crenças (LIU; MAN, 2005), redes probabilísticas (LOUZADA; ARA, 2012) ou redes causais (BOBBIO et al., 2001), são baseadas nas teorias da probabilidade e dos grafos (PEARL, 1988). Nessas redes, variáveis aleatórias são visualmente representadas por vértices, também denominados por nós; a existência de um arco direcionado entre vértices indica a dependência probabilística direta entre as variáveis aleatórias conectadas, de modo a formar um grafo acíclico direcionado. Vale ressaltar que em uma rede bayesiana por ser um *I-map*, mapa de independência, sendo que ausência de arco entre duas variáveis implica na independência condicional entre elas (NADKARNI; SHENOY, 2001). Então, os arcos são utilizados para indicar a (in)dependência condicional entre pares de variáveis, a qual é baseada na noção de *d-separação* (do inglês, *directional - separation*, em tradução literal, separação direcional), a qual permite a leitura de estruturas de independência (KOLLER; FRIEDMAN, 2009). Além disso, esse mesmo conceito pode permitir a interpretação causal entre variáveis (PEARL, 2000). Neste contexto, a estrutura de uma RB pode ser descrita por especialistas ou pode ser estimada por meio de algoritmos de aprendizado.

Um dos desafios no uso de RB é o processo de estimação da estrutura geral da rede para dados reais, uma vez que pode ser inviável a descrição da estrutura por meio de especialistas. Os algoritmos de aprendizado da estrutura presentes na literatura podem ser classificados em três abordagens (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019):

i) baseadas em pontuação, que se baseiam em um espaço de hipóteses quanto à estrutura da rede e em uma métrica de pontuação, em outras palavras, uma função objetivo que deve ser minimizada no sentido de encontrar a estrutura que melhor se ajuste aos dados, ii) baseadas

em restrições, que buscam por uma estrutura que melhor explique as dependências e independências encontradas nos dados, iii) híbridas, que utilizam as duas abordagens apresentadas anteriormente.

Deste modo, vários métodos de estimação de estrutura têm sido propostos por diversos autores (COOPER; HERSKOVITS, 1992; SPIRTES et al., 2000; COLOMBO; MAATHUIS, 2014; SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019). Se por um lado métodos baseados em pontuação podem resultar em redes densas, por outro lado métodos baseados em restrições podem gerar RB esparsas. As metodologias que unem ambas as naturezas se propõem a superar as limitações das técnicas individualmente (GASSE; AUSSEM; ELGHAZEL, 2012). Neste artigo, é proposto um método híbrido de estimação de estrutura, denominado *scoring and restrict*. Este método é referente a uma abordagem híbrida que se baseia na junção de duas abordagens diferentes de estimação. Primeiramente, encontra-se a estrutura que maximiza uma função geral de ajuste aos dados. Posteriormente, insere-se a informação referente à variável de interesse de forma restritiva, sendo este último que se baseiando-se em testes de independência condicional. O resultado dessa incorporação é um modelo gráfico equilibrado entre redes de conexões esparsas e redes densas, proporcionando muitas vezes, inclusive, aumento da capacidade preditiva do modelo. Tal conclusão é respaldada na investigação feita por meio de simulação e na utilização de dados de agricultura para a predição da porcentagem de falha em plantações de cana-de-açúcar.

Este artigo começa apresentando os principais fundamentos de Redes Bayesianas, bem como uma breve descrição dos métodos tradicionais de estimação de estrutura. Em seguida, é detalhado o método de estimação proposto e um estudo de simulação é conduzido. Por fim, são descritos e discutidos os resultados finais da aplicação, e exibidas as considerações finais.

Redes Bayesianas

As Redes Bayesianas foram introduzidas por Judea Pearl (PEARL, 1988) no fim dos anos 80 e são um tipo de modelo gráfico probabilístico composto por dois elementos essenciais. O primeiro componente refere-se à estrutura gráfica da rede, G, a qual é definida como sendo um Grafo Acíclico Direcionado (do inglês, DAG - *Directed Acyclic Graph*), pois não permite ciclos e arcos não direcionados. O DAG, por sua vez, também é composto por dois elementos, que são os nós e os arcos direcionados. Eles são as representações visuais dos elementos do segundo componente das redes, a distribuição conjunta de probabilidade P. A distribuição de probabilidade é obtida com respeito as variáveis aleatórias as quais podem ser de natureza numérica - contínua ou discreta - ou categórica. Tradicionalmente, as RBs que consideram apenas variáveis categóricas ou discretizadas são chamadas de Redes Bayesianas discretas, ou simplesmente Redes Bayesianas, uma vez que o tratamento de variáveis contínuas é comumente realizado através da suposição de normalidade multivariada, sendo chamadas de Redes Gaussianas (GEIGER; HECKERMAN, 1994).

Dada uma distribuição de probabilidade conjunta $P(X)$ sobre um conjunto de variáveis aleatórias $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ e um conjunto de arcos A, um DAG $G = (\mathbf{X}, A)$ é uma RB se e somente se $L \perp\!\!\!\perp_G M \mid N \Rightarrow L \perp\!\!\!\perp_P M \mid N$, onde $L, M, N \subseteq \mathbf{X}$ são conjuntos disjuntos. Dessa forma, a separação condicional em G implica em independência condicional na distribuição de probabilidade P. Essa definição é obtida por meio da noção de *d-separação* a qual permite inferir propriedades de independência de uma distribuição de probabilidade P fatorada de um grafo G, com a análise simples da conectividade da estrutura (KOLLER; FRIEDMAN, 2009). Maiores detalhes sobre estas propriedades podem ser consultadas em Korb e Nicholson (2010) e Scutari e Denis (2014).

Desta forma, o modelo satisfaz a condição de que cada um dos elementos do vetor aleatório $\mathbf{X}$ é independente do conjunto de seus não descendentes, condicionados ao conjunto de seus pais. Esta condição é também conhecida como condição de Markov. Isso permite que a distribuição global P, seja fatorada em distribuições locais de probabilidade. Em outras palavras, a

distribuição conjunta de probabilidade é o decomposta por meio de funções de probabilidade condicionais, como exibe a Equação (1).

$$P(\mathbf{X}) = \prod_{i=1}^{d} P\left(X_i | \mathbf{pa}_G(X_i)\right), \tag{1}$$

sendo $\mathbf{pa}_G(X_i)$ o conjunto de pais de X_i, o qual refere-se a existência de arcos direcionados das variáveis do conjunto $\mathbf{pa}_G(X_i)$ para X_i, sendo que sua cardinalidade pode variar para cada um dos elementos de $\mathbf{X}$. Além disso, os componentes desse conjunto de pais estão condicionados a estrutura gráfica do DAG G, o qual deve decodificar a função de probabilidade P (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019).

Para cada uma dessas variáveis, necessariamente, deve existir $P(X_i|\mathbf{pa}_G(X_i))$ que é chamada de distribuição local (NIELSEN; JENSEN, 2009). Essas distribuições são basicamente tabelas (ou densidades, para o caso contínuo) condicionais de cada uma delas (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019). Contudo, o conjunto de pais de quaisquer variáveis pode ser vazio, sendo a variável apenas pai de outra variável ou, ainda, independentemente do restante da estrutura.

Assim, o aprendizado de Redes Bayesiana busca por modelos que melhor representam a maneira com que as variáveis se conectam entre si e como tais conexões formam mapas de dependência entre seus nós (KOLLER; FRIEDMAN, 2009). Como as redes são compostas em dois elementos, sua estimação é dividida tradicionalmente em duas tarefas principais: a estimação da estrutura G e dos parâmetros de $P(X)$. A estrutura de uma RB pode ser descrita por especialistas e, nesse caso, a estimação será direcionada apenas aos parâmetros; ou podem ser obtidas por meio de algoritmos de aprendizado, como será apresentado a seguir.

Aprendizado de estruturas em Redes Bayesianas

No aprendizado das estruturas, direcionado a estimar as conexões dentre as variáveis para um DAG, todos os nós são considerados irrestritos. Desta forma, não é necessário especificar uma ou mais variáveis respostas, todas as variáveis da rede são consideradas aleatórias. Em outras palavras, não é necessário especificar uma única variável resposta e, caso esta exista, será considerada qualquer outra variável, não possuindo nenhuma regra preestabelecida de conexão, como ocorre com alguns casos especiais de Redes Bayesianas para classificação, como, por exemplo, para o classificador *Naïve Bayes* (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997; BIELZA; LARRAÑAGA, 2014).

Neste contexto, as heurísticas utilizadas para cumprir a tarefa de estimação de estrutura são inúmeras (BERETTA et al., 2018), porém podem ser classificadas em três principais grupos: o primeiro se baseia na maximização de uma função de ajuste na busca da melhor estrutura de relação entre variáveis, chamado de aprendizagem baseada em métricas (*score-based*, *score and search*, ou *scoring*) (SU et al., 2012.), o segundo é baseado em testes de independência condicional, chamado de aprendizagem de restrição (*constraint-based*) (COLOMBO; MAATHUIS, 2014) e o terceiro refere-se a combinação dos aspectos de busca e maximização de funções e testes de independência condicional (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019).

Estimação baseada em métricas

O procedimento do aprendizado de um algoritmo baseado em métricas é uma busca por conexões, seguida do cálculo de uma função objetivo que mede o grau ajuste da rede, que deve ser escolhida previamente (BERETTA et al., 2018), bem como posterior comparação com o melhor resultado até então obtido. A repetição desses passos visa encontrar a melhor estrutura $g*$ no espaço $\mathcal{G}$ a qual maximize a função objetivo $f(g, D)$, tal que,

$$g* = \arg\max_{g \in \mathcal{G}} f(g, D),$$

sendo que D é o conjunto de dados disponíveis para estimação (CAMPOS, 2006).

Assim, o algoritmo realiza uma busca via *greedy search* no espaço de possíveis grafos $\mathcal{G}$, a fim de encontrar uma estrutura que não cause perda da medida em relação a opção anterior. Um algoritmo amplamente utilizado é o *Hill-Climbing* (GÁMEZ; MATEO; PUERTA, 2011) porém, uma versão otimizada desse algoritmo chamada *Tabu Search* tenta evitar máximos locais limitando os locais que já foram testados anteriormente, na lista *tabu*, que não pode ser revisitada (RUSSELL; NORVIG, 2010). Além disso, quando este método encontra um máximo, realiza mais algumas iterações na tentativa de garantir que tenha atingido um máximo global (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019).

Dentre as métricas mais comumente utilizadas, Cooper e Herskovits (1992) descrevem um algoritmo de busca associado a função objetivo *K2* a ser maximizada. O objetivo final é encontrar uma estrutura que seja razoável e obtenha o melhor valor dentre todos os modelos possíveis (ABELLÁN et al., 2006; LERNER; MALKA, 2011).

Por ser uma medida de ajuste de toda a rede, é calculada com respeito ao conjunto de observações para cada variável contida em $\mathbf{X}$. O cálculo é condicionado a estrutura do grafo e a probabilidade da estrutura G é dada por $P(G)$, sendo expressa por,

$$f_{K_2}(D|G) = log(P(G)) + \sum_{i=1}^{d+1} \left(\sum_{j=1}^{q_i} \left(log \left(\frac{(c_i - 1)!}{(N_{ij} + c_i - 1)!} \right) + \sum_{k=1}^{c_i} log\left(N_{ijk}!\right) \right) \right),$$

sendo que, o número de classes da variável X_i é dado por c_i, o número de combinações das variáveis em $\mathbf{pa}_G(X_i)$ de X_i é q_i, N_{ijk} é o número de observações na base de dados na qual a variável X_i recebe o valor x_{jk} e $N_{ij} = \sum_{k=1}^{ri} N_{ijk}$. Essa medida é baseada em diversas suposições como as de multinomialidade e independência de parâmetros (CAMPOS, 2006).

Neste artigo, o método de estimação baseado na métrica *K2* realizado através da busca via *Tabu Search* será denominado, simplesmente, como algoritmo *K2*.

Estimação baseada em testes

Esse grupo de algoritmos se baseia em testes de independência condicional para verificar a existência de arcos. Nesse contexto, quando se constata a independência entre duas variáveis, o arco entre aquelas variáveis é retirado (ZANG et al., 2012).

Dentre os algoritmos mais comuns tem-se o algoritmo PC (SPIRTES et al., 2000) sendo modernizado por Colombo e Maathuis (2014), também conhecido como *PC-stable*, uma vez que não realiza o pressuposto de ordenação de variáveis.

O algoritmo *PC-Stable* inicia a busca com um grafo completo não-direcionado, isto é, inicialmente todos os nós são conectados aos outros por arestas e, durante sua execução, se ocupa em avaliar localmente a necessidade de arcos conectando cada par de nós (NIELSEN; JENSEN, 2009). Essa verificação é feita por meio de testes de independência condicional baseado em uma métrica pré-definida.

Neste artigo, a estatística definida para os testes de independência condicional foi a Informação Mútua Condicional (CAMPOS, 2006), dada pela Expressão (2).

$$I(X_i, X_j|Z) = \sum_{z \in Z} \left(P(z) \sum_{x_j \in X_j} \sum_{x_i \in X_i} P(x_i, x_j|z) \log \left(\frac{P(x_i, x_j|z)}{P(x_i|z)P(x_j|z)} \right) \right), \quad (2)$$

sendo que para cada par de variáveis x_i e x_j é considerado um conjunto condicionante Z.

De acordo com Nielsen e Jensen (2009), métodos baseados em testes apresentam propriedades importantes para a construção das redes e seus atributos causais: i) se a base de dados for um conjunto fiel de uma Rede Bayesiana, então o grafo resultante dessa sequência de passos é o esqueleto da rede; ii) as independências condicionais encontradas pelo algoritmo são suficientes para determinar as estruturas-v, determinando assim, as direções dos arcos.

Uma vez que é baseado em testes, essa categoria de algoritmos de aprendizado de estrutura pode possuir dificuldade em adicionar ou até mesmo direcionar arcos. Gerando estruturas não-conectadas ou não-direcionadas, as quais não podem ser consideradas em um DAG.

Estimação híbrida

Por sua vez, os algoritmos híbridos combinam características de ambos os grupos descritos acima, e têm o objetivo de restringir e maximizar algum tipo de métrica pré-definida, ou vice-versa (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019). Mesmo que ainda pouco utilizados em comparação com as metodologias anteriores, a interação entre essas metodologias pode permitir a minimização dos efeitos das limitações de cada algoritmo. Existem algoritmos híbridos já preestabelecidos que se utilizam dos algoritmos PC e Hill-Climbing, como o *HPC - Hibrid PC* e o *Max-Min Hill-Climbing* (GASSE; AUSSEM; ELGHAZEL, 2012). Ainda, outras formas de utilização que trabalham com a combinação dos métodos utilizando a informação obtida em um dos métodos como inicialização para o outro. Geralmente, a forma híbrida de combinação entre variáveis se baseia na busca pela melhor estrutura de conexão seguida da verificação de independência condicional entre os arcos.

Comentários gerais

É importante salientar que o número de parâmetros a serem estimados na distribuição conjunta é exponencial em termos do número de variáveis (NEAPOLITAN et al., 2004), do número de classes ou categorias de cada uma delas, além da quantidade de conexões da rede, e isso faz com que o problema de estimação seja de complexidade *NP-hard*. Contudo, Bielza e Larrañaga (2014) afirmam que para predição é necessário saber apenas sobre o conjunto que constitui a Cobertura de Markov da variável de interesse, sendo assim, a complexidade pode ser reduzida. Além disso, o foco é o aprendizado de estrutura das Redes Bayesianas, contudo a estimação dos parâmetros é a tarefa posterior ao acabamento da arquitetura do grafo. Os parâmetros também podem ser estimados de diversas maneiras nesse texto, por se tratar de redes discretas, serão estimados via o método tradicional Multinomial-Dirichlet entre estados das variáveis e suas conexões (NAGARAJAN; SCUTARI; LÈBRE, 2013).

Método híbrido *scoring and restrict*

Neste artigo considera-se um novo método de estimação híbrida baseado na existência de uma variável específica a ser predita, visto que, em geral, os processos de estimação de estrutura em Redes Bayesianas consideram que todas as variáveis possuem um mesmo nível de interesse e não focam suas restrições em uma variável especial.

Sendo assim, primeiramente um método *score-based* via algoritmo *K2* é aplicado na base de dados e, a partir dessa estrutura baseada na maximização da função objetivo $K2$, verificam-se as conexões das covariáveis com a variável específica a ser predita, que são parte fundamental da sua Cobertura de Markov. Assim, o método proposto possui um caráter mais preditivo do que os métodos citados anteriormente. De uma forma geral, essas conexões são armazenadas em uma lista que especifica a incidência do arco, ou seja, indicadores de direção das covariáveis para a variável a ser predita.

Esse conjunto de conexões direcionadas são inseridas na aplicação do segundo método, um método baseado em testes. Neste caso, utilizamos o *PC-stable* que é baseado em testes de

independência condicional. Essa lista é chamada de *whitelist* e a sua utilização faz com que o algoritmo deixe de realizar o cálculo de relação entre as variáveis que compõem cada um dos arcos desse grupo, já que essas são estabelecidas antes desse aprendizado. O restante das conexões entre variáveis continuará a ser estimada conforme o procedimento do método.

Por fim, a metodologia híbrida recebe o nome de *scoring and restrict*, nessa ordem, pois primeiramente, estima-se uma estrutura de rede maximizando uma métrica (*K2*) e posteriormente, se utiliza das conexões encontradas no resultado anterior relativas à variável de interesse, como um grupo de restrições minimizando o espaço de busca de estruturas para o algoritmo baseado em testes de independência condicional (*PC*). Essa nova proposta será denotada nesse artigo como *K2+PC*.

No Algoritmo 1 o pseudocódigo do método *scoring and restrict* é apresentado. Como parâmetros de entrada, uma vez que pode ser utilizado através da heurística *tabu-search*, o algoritmo requer: i) o comprimento da lista de tabu usada na função tabu L_1, ii) número de iterações tabu que podem ser executadas sem melhorar a pontuação da rede L_2, iii) conjunto de dados D, e iv) teste de independência ϕ.

input : variável de resposta $y \in X$, conjunto de dados D, comprimento da lista de tabu L_1; número de iterações tabu que podem ser executadas sem melhorar a pontuação da rede L_2; teste de independência condicional ϕ.
output: Para cada nó $X \in \mathbf{X}$, uma lista de pais $pa(X)$
1 $S = K2(y, L_1, L_2, D)$;
2 $whitelist = \{(u, y) \in S\}$;
3 **return** $PC(whitelist, y, D, \phi)$;

Algoritmo 1: Pseudocódigo do método proposto *K2+PC*.

Como saída, o algoritmo retorna a lista de pais $pa(x)$ para cada variável aleatória $X \in \mathbf{X}$. Na linha 1 do algoritmo, executamos o algoritmo K2, conforme apresentado em Russel e Norvig (2010), de modo que S representa o conjunto de arcos retornado pelo algoritmo K2. Na linha 2, define-se um conjunto *whitelist* formado por arcos $(u, y) \in S$ incidentes sobre a variável de resposta $y \in X$. Por fim, na linha 3, executamos o algoritmo PC conforme apresentado em Scutari (2010), porém com as seguintes modificações: i) em vez de iniciar sua execução com um grafo completo sobre X, o algoritmo PC é iniciado com um subgrafo completo induzido $G[X - \{y\}]$ unido ao dígrafo $\bar{G} = (X, whitelist)$, e ii) arcos presentes em *whitelist* não são considerados para teste de independência condicional e, então, não são passíveis de exclusão.

Estudo de simulação

O experimento de simulação é projetado para responder as seguintes perguntas: i) O nosso método *scoring and restrict* é eficaz para estimação de estrutura? ii) Como nosso método se compara com o tradicional do algoritmo *K2* ? Para responder essas perguntas, utilizamos conjuntos de dados sintéticos, além de *baselines* para comparação. A seguir, tais configurações são descritas.

Conjunto de dados artificiais

A base de dados corresponde a três quantidades distintas de observações $\{100, 500, 1000\}$, bem como são compostas por 5 variáveis explicativas $\mathbf{X} = \{X_1, X_2, \ldots, X_5\}$ e uma variável resposta Y. Suas dependências probabilísticas foram simuladas através das relações apresentadas a seguir:

$$
\begin{array}{rll}
Y \mid X_3, X_4, X_5 \sim \text{Gamma}(\mu, \sigma) & \log(\mu) = 1.2\,X_3 + 2.5\,X_4 - 0.2\,X_4^2 + 1.5\,X_5 & \sigma = 0.65 \\
X_3 \mid X_1, X_2 \sim \text{Beta}(\nu, \phi) & \text{logit}(\nu) = 0.1\,X_1 + 0.05\,X_2 & \phi = 0.7 \\
X_1 \sim \text{Bernoulli}(\pi) & \pi = 0.3 & \\
X_2 \sim \text{Poisson}(\lambda) & \lambda = 25 & \\
X_4 \sim \text{Normal}(\eta, \tau) & \eta = 0.5 & \tau = 0.1 \\
X_5 \sim \text{Bernoulli}(\theta) & \theta = 0.65 &
\end{array}
$$

O grafo teórico dos dados simulados possui a estrutura apresentada na Figura 1; Y tem como pais X_3, X_4, X_5, enquanto X_3, por sua vez, é filha de X_1 e X_2. Esta forma de controle da natureza de geração das variáveis utilizadas visa entender o padrão de desempenho dos métodos da estimação de estrutura utilizados nesse estudo.

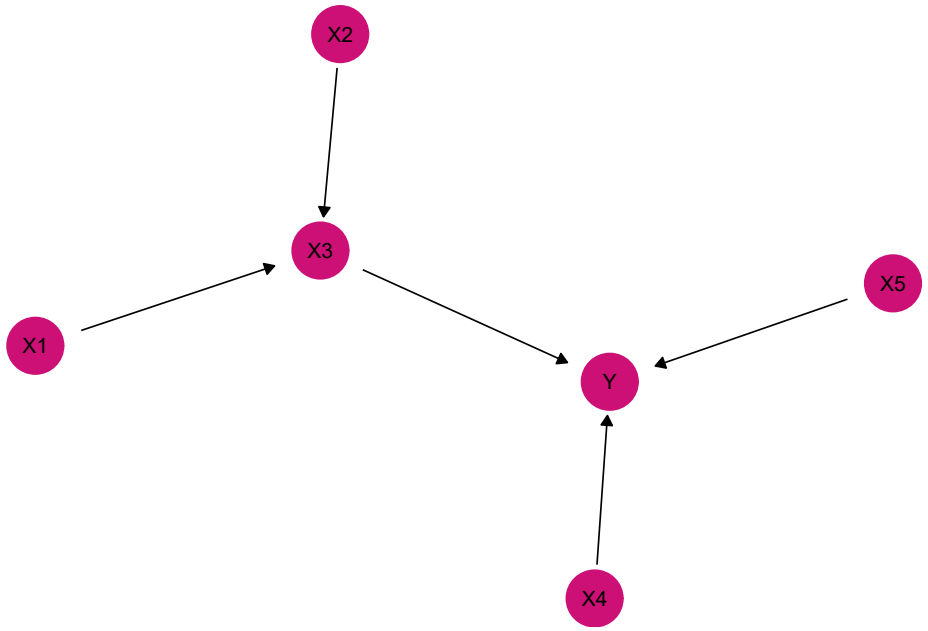

Figura 1: Grafo da estrutura teórica simulada.

Dada a estrutura teórica gráfica, o número de parâmetros para cada configuração aumenta exponencialmente conforme discutido na seção sobre estimação. A quantidade de parâmetros para o grafo da Figura 1 é dado por $2 + L_{X_2}\,(1 + L_{X3}L_{X1}) + 2L_{X4}\,(L_Y L_{X3} + 0,5L_{X5})$, com L_{Xi} e $i = 1, \ldots, 5$ as classes de discretização das variáveis explicativas e L_Y as classes da variável resposta. Detalhes em Korb e Nicholson (2010), pág. 33. Desta forma, sendo o número mínimo de parâmetros é igual a 32 para todas as variáveis binárias e o número máximo igual a 182 para $L_{Xi} = 4$ e $L_Y = 3$, número máximo de classes nos procedimentos de discretização utilizados neste artigo.

Metodologia de Avaliação

O método proposto requer que as variáveis aleatórias sejam discretas e, então, a base de dados deve possuir apenas variáveis categóricas. Para atender essa restrição, quando necessário, foi aplicado um método de discretização de variáveis numéricas em classes de mesma frequência. A escolha da quantidade de classes é um dos fatores de mudança na configuração da rede e portanto, do tempo computacional necessário para cálculo das estimações de estrutura e parâmetros.

Desta forma, a metodologia descrita na Seção Método Híbrido é integralmente aplicada em cada um das bases de dados para configurações de 2 ou 3 classes de discretização da variável

resposta e 2, 3 ou 4 classes de discretização para as covariáveis que possuem natureza numérica e precisam, portanto, serem categorizadas.

Assim, as análises das alterações de performance preditiva da rede estimada devem ser verificadas por meio da quantidade de classes escolhidas via o processo de discretização, bem como por meio da interpretação da estrutura estimada, levando-se em consideração a escolha de modelos parcimoniosos.

A estimação das redes é realizada através da base de dados completa por meio do procedimento de amostras *bootstrap*, com 100 replicações e 10 reinícios para *Tabu-Search*. A rede estimada considera uma incidência de arcos e direcionamentos maiores que 50% para as redes geradas via amostras *bootstrap* (FRIEDMAN; GOLDSZMIDT; WYNER, 2013), tanto na estimação de rede pelo método *K2* quanto por meio do algoritmo *K2+PC*. A estimação de parâmetros foi feita por meio da metodologia de *hold-out* repetido, considerando 100 repetições de amostras distintas 70/30 - a média desse *loop* é utilizada como medida de performance comparativa. De modo que 30% da amostra de cada repetição foi utilizada para cálculo das medidas preditivas. As métricas de avaliação utilizadas são listadas a seguir:

- **Acurácia**: a fração de predições corretas do modelo, dada por:

$$ACC = \frac{\sum_{i=1}^{k} C_{ii}}{\sum_{i,j=1}^{k} C_{ij}},$$

  sendo k é o número de classes as quais os resultados podem ser observados, $\sum_{i=1}^{k} C_{ii}$ é a somas das predições corretas e $\sum_{i,j=1}^{k} C_{ij}$ é o número total de observações.

- **Coeficiente de Correlação de Matthew**: é utilizada para verificar a classificação geral do modelo e é interpretada similarmente ao coeficiente de correlação de Pearson (LOUZADA; ARA, 2012), visto que é uma generalização dessa medida (GORODKIN, 2004), quanto mais próxima a 1, mais ajustada está a predição aos resultados observados. Sua forma multiclasse é dada pela expressão:

$$MCC = \frac{\sum_{k,l,m=1}^{N} C_{kk}C_{ml} - C_{lk}C_{km}}{\sqrt{S_{k1}}\sqrt{S_{k2}}},$$

  com $S_{k1} = \sum_{k=1}^{N} \left[\left(\sum_{l=1}^{N} C_{lk} \right) \left(\sum_{f,g=1 f \neq k}^{N} C_{gf} \right) \right]$ e
  $S_{k2} = \sum_{k=1}^{N} \left[\left(\sum_{l=1}^{N} C_{kl} \right) \left(\sum_{f,g=1 f \neq k}^{N} C_{fg} \right) \right]$.

- **Spherical Payoff**: em tradução literal do inglês, recompensa esférica, é a medida que dimensiona a capacidade do modelo em atribuir probabilidades semelhantes as reais para cada uma das categorias. É uma medida considerada de ajuste fino e varia de 0 a 1 onde 1 é a melhor performance do modelo, é definida como:

$$SP = MOAC \frac{P_c}{\sqrt{\sum_{j=1}^{N} P_j^2}},$$

  sendo $MOAC$ (sigla para *mean over all cases*) é a média entre todos os casos, P_c é a probabilidade atribuída a real classe do indivíduo, P_j é a probabilidade de cada estado da variável categórica e n é o número de estados da variável categórica (MARCOT, 2012).

Após o aprendizado da estrutura de Redes Bayesianas discretas, as tabelas de probabilidade condicional foram estimadas pelo método tradicional de conjugação Dirichlet-multinomial (RUZ; ARAYA-DÍAZ, 2018).

Ambiente computacional

Todas as análises foram realizadas através de um computador pessoal Intel(R) Core(TM) i7-8565U - CPU 1.80GHz 1.99GHz, memória RAM de 16GB e sistema operacional *Windows* 10. Por meio do software `R` (R CORE TEAM, 2018), versão 4.0.2, foi utilizado por meio dos pacotes `bnlearn` (v.4.5) (SCUTARI, 2010), utilizado para ajuste das redes, `infotheo` (v.1.2.0) (MEYER, 2014) e `network` (v.1.16.0) (BUTTS, 2008). As configurações se baseiam no padrão dos métodos implementados via `bnlearn`, o qual já possui a implementação dos algoritmos *K2* e *PC-Stable*, porém não possui a implementação do algoritmo *K2+PC*. O número de comprimento da lista tabu e do número iterações tabu que podem ser executadas sem melhorar a melhor pontuação da rede foi fixado em 10, valor padrão utilizado na solução de problemas simples (TSUBAKITANI; EVANS, 1998; SCUTARI; DENIS, 2014). O teste de independência condicional é baseado no teste de informação mútua condicional.

Resultados da Simulação

A Tabela 1 contém os resultados da simulação para cada um dos 36 modelos gerados através das diferentes metodologias: número de classes de discretização de X e Y e quantidade de registros. Os valores estão apresentados multiplicados por 100, tanto para média quanto para o desvio padrão na estrutura de {média $\pm$ desvio padrão}; a ausência de valores indica que o algoritmo não foi capaz de gerar uma rede DAG final, sendo incapaz de realizar o cálculo dos parâmetros.

Naturalmente, quando a quantidade de classes a serem preditas é menor, os valores das medidas preditivas são mais elevados, portanto, a comparação deve levar em consideração a comparação da capacidade preditiva apenas entre os métodos para mesma configuração de rede. No geral, existe um ganho médio de 4% de acurácia na utilização da metodologia K2+PC ao invés da metodologia K2, da mesma forma o coeficiente de correlação de Mathew tem acréscimo médio de $3,17\%$ e o *Spherical Payoff* tem aumento médio de $2,25\%$, sugerindo maior poder preditivo quando o método híbrido de estimação de estrutura é utilizado. Essa conclusão é reforçada pela análise gráfica das estruturas estimadas.

Desta forma, as estruturadas estimadas são apresentadas nas Figuras de 2 a 5. Cada uma faz referência ao método e a quantidade de classes de discretização da variável resposta. Assim, as quais possuem a quantidade de observações dada pelas colunas, então a primeira coluna corresponde ao tamanho 100, a segunda coluna ao tamanho 500 e terceira coluna ao tamanho 1000. Semelhantemente, as linhas correspondem a quantidade de classes as quais as variáveis explicativas foram discretizadas, de forma que a linha um corresponde a 2 classes, a linha dois corresponde a 3 classes e a linha três corresponde a 4 classes.

As Figuras 2 e 3, nas quais Y é dicotomizado, correspondem respectivamente às metodologias *K2* e *K2+PC*. Na primeira, observa-se grafos bastante conectados quando comparados a segunda, bem como há uma concordância da estrutura estimada com a estrutura gerada em apenas 3 dos 9 grafos da Figura 2, itens (c), (f) e (g), para a Figura 3 têm-se 4 dos 9, itens (b), (c), (e) e (i). Analogamente para as Figuras 4 e 5, as quais correspondem ao número de categorias de Y igual a 3 e as metodologias *K2* e *K2+PC* respectivamente, nenhuma das estruturas estimadas da Figura 4 obtiveram acertos com relação a estrutura gerada. Contudo, para a Figura 5 têm-se conformidade com a estrutura teórica para 5 dos 9 grafos estimados, itens (b), (c), (e), (f) e (h). Para um menor tamanho amostral, primeira coluna das figuras, nota-se uma confusão de ambas as metodologias, tanto em relação à conectividade quanto em relação ao direcionamento dos arcos.

Além disso, nota-se que o algoritmo *K2* resulta em uma maior conectividade entre as variáveis, gerando um maior número de arcos. A distribuição do número de arcos de ambas as metodologias é expressa pela Figura 6. Assim, nota-se que a metodologia *K2+PC* é mais parcimoniosa e sem perda de capacidade preditiva. O alto número de conexões do algoritmo

K2 é bem expresso pela Figura 4 e pela Figura 2, esta última para o caso de um tamanho amostral pequeno (primeira coluna). Este fato de inflacionamento dos arcos pode prejudicar a interpretação da estrutura estimada para problemas reais.

Em termos de tempo computacional, para todas as iterações de amostragem *bootstrap* e ajuste do modelo, obteve-se um mínimo de 0,61 segundos e máximo de 2,04 segundos para o ajuste da rede pelo método *K2*. Quando o método *PC* é inserido, o tempo aumenta 57%, em média, passando para um tempo mínimo de 1,05 segundos e máximo de 2,89 segundos.

Tabela 1: Resultados dos dados simulados (valores das medidas preditivas multiplicados por 100).

Número de Registros	Classes Y	Classes X	Algoritmo	Arcos	Acurácia	Coeficiente Correlacão de Mathews	Spherical Payoff
100	2	2	K2	9	61,0 ± 7,1	63,7 ± 7,0	73,2 ± 3,5
			K2+PC	4	79,5 ± 6,8	80,6 ± 6,7	82,6 ± 4,2
		3	K2	13	77,9 ± 6,5	78,9 ± 6,2	82,3 ± 3,8
			K2+PC	3	77,4 ± 5,8	78,4 ± 5,6	82,3 ± 3,8
		4	K2	11	60,0 ± 6,3	62,7 ± 6,6	73,0 ± 3,3
			K2+PC	3	80,1 ± 6,6	80,7 ± 6,3	83,0 ± 5,0
	3	2	K2	10	-	-	-
			K2+PC	4	63,4 ± 8,1	73,4 ± 5,8	71,6 ± 5,1
		3	K2	12	59,9 ± 7,9	70,9 ± 5,8	70,2 ± 5,0
			K2+PC	3	60,1 ± 7,8	71,1 ± 5,8	70,2 ± 5,0
		4	K2	14	48,5 ± 7,9	63,2 ± 5,7	61,7 ± 4,4
			K2+PC	2	48,7 ± 7,9	63,3 ± 5,7	61,7 ± 4,4
500	2	2	K2	6	79,6 ± 2,7	80,1 ± 2,6	84,9 ± 1,8
			K2+PC	5	79,6 ± 2,7	80,1 ± 2,6	84,9 ± 1,8
		3	K2	5	80,1 ± 2,6	80,3 ± 2,6	83,7 ± 1,7
			K2+PC	5	78,9 ± 2,8	79,2 ± 2,8	84,9 ± 1,7
		4	K2	9	78,8 ± 2,3	79,4 ± 2,3	84,6 ± 1,6
			K2+PC	6	81,0 ± 2,8	81,1 ± 2,7	85,6 ± 1,8
	3	2	K2	5	64,2 ± 3,4	73,4 ± 2,5	72,6 ± 1,9
			K2+PC	4	**64,2 ± 3,5**	**73,4 ± 2,6**	**72,6 ± 1,9**
		3	K2	7	60,1 ± 3,3	70,9 ± 2,3	72,2 ± 1,7
			K2+PC	5	**65,3 ± 3,7**	**74,2 ± 2,7**	**73,7 ± 2,0**
		4	K2	10	65,1 ± 3,2	74,1 ± 2,4	73,2 ± 1,7
			K2+PC	5	**66,1 ± 3,0**	**74,7 ± 2,2**	**74,6 ± 2,0**
1000	2	2	K2	5	77,3 ± 1,7	78,8 ± 1,5	82,9 ± 0,9
			K2+PC	5	78,8 ± 1,9	79,1 ± 1,8	84,3 ± 1,1
		3	K2	5	82,5 ± 1,9	82,7 ± 1,9	85,9 ± 1,2
			K2+PC	4	82,5 ± 1,7	82,7 ± 1,7	87,3 ± 1,1
		4	K2	6	83,1 ± 1,8	83,6 ± 1,7	86,3 ± 1,2
			K2+PC	5	84,9 ± 1,5	85,0 ± 1,5	87,8 ± 1,1
	3	2	K2	5	61,9 ± 2,2	71,9 ± 1,8	73,0 ± 1,3
			K2+PC	5	**61,9 ± 2,3**	**71,9 ± 1,8**	**73,0 ± 1,3**
		3	K2	7	69,9 ± 2,1	77,6 ± 1,5	76,4 ± 1,3
			K2+PC	5	**69,8 ± 2,1**	**77,6 ± 1,6**	**76,4 ± 1,3**
		4	K2	7	67,4 ± 2,3	75,9 ± 1,7	74,8 ± 1,2
			K2+PC	6	71,0 ± 2,3	78,3 ± 1,7	77,4 ± 1,3

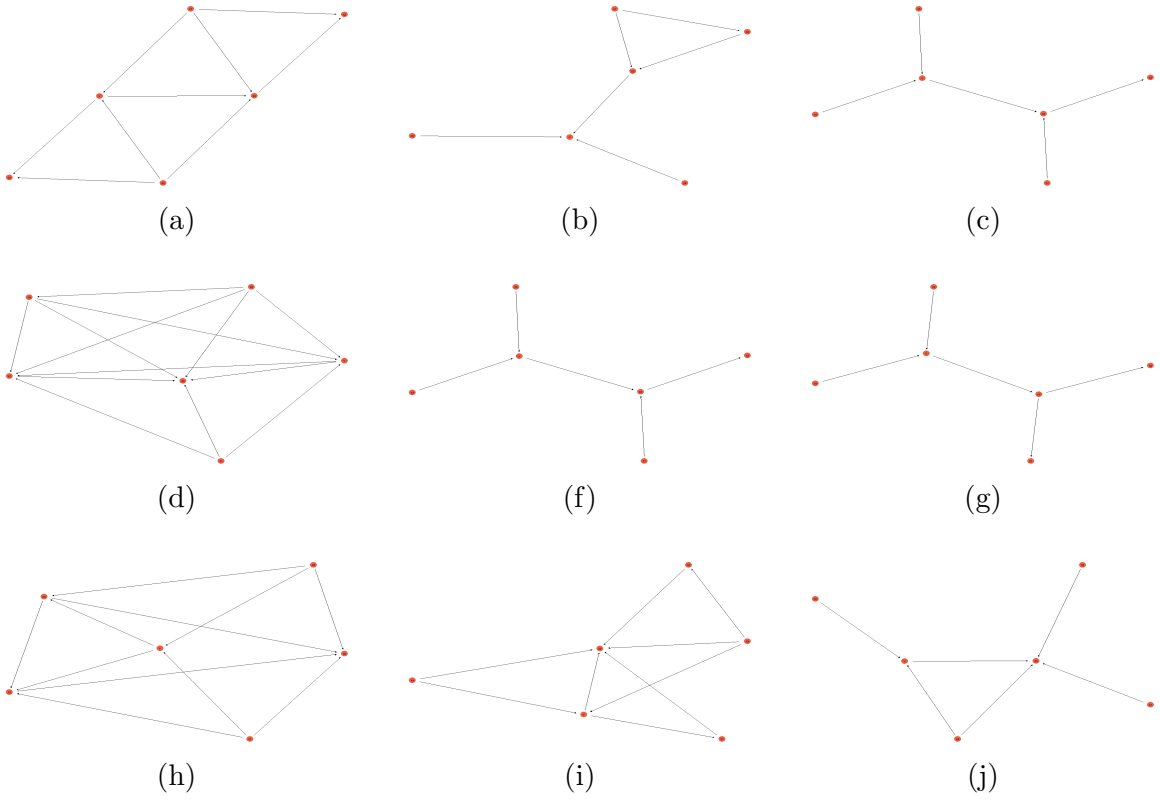

Figura 2: Estruturas obtidas por meio do método K2, para número de classes da variável resposta igual a 2.

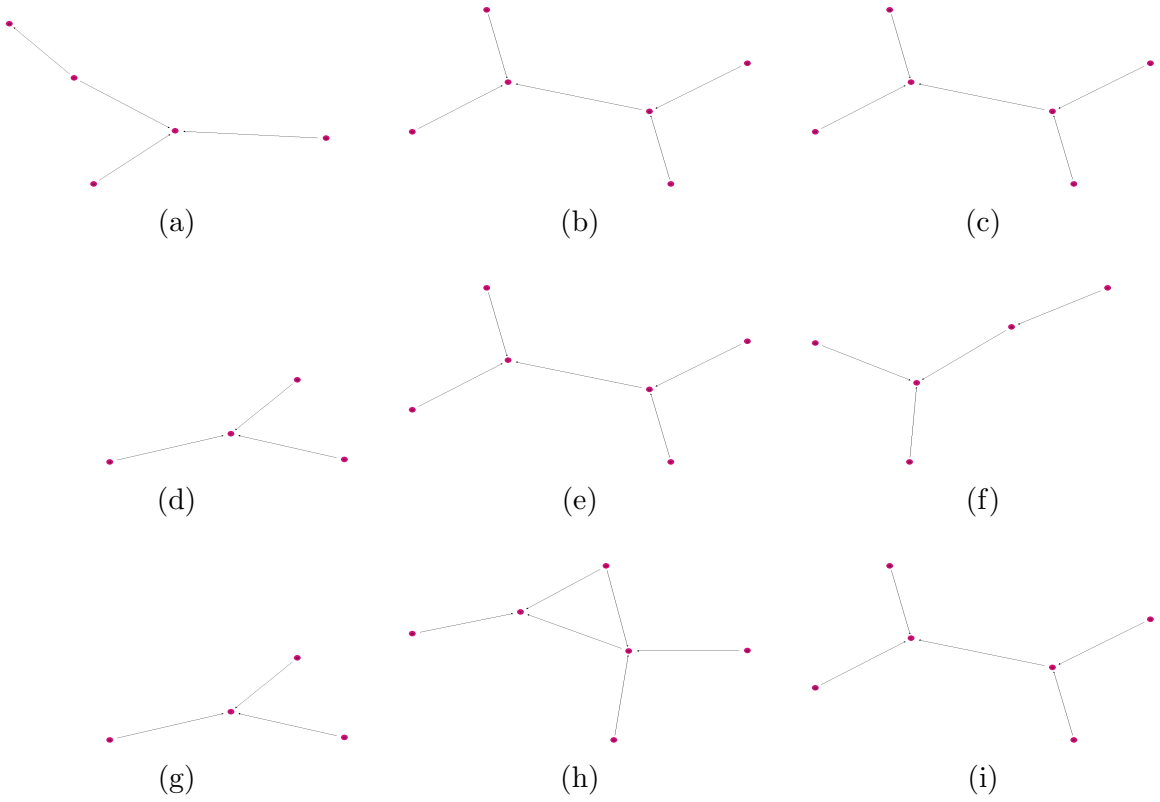

Figura 3: Estruturas obtidas por meio do método K2+PC, para número de classes da variável resposta igual a 2.

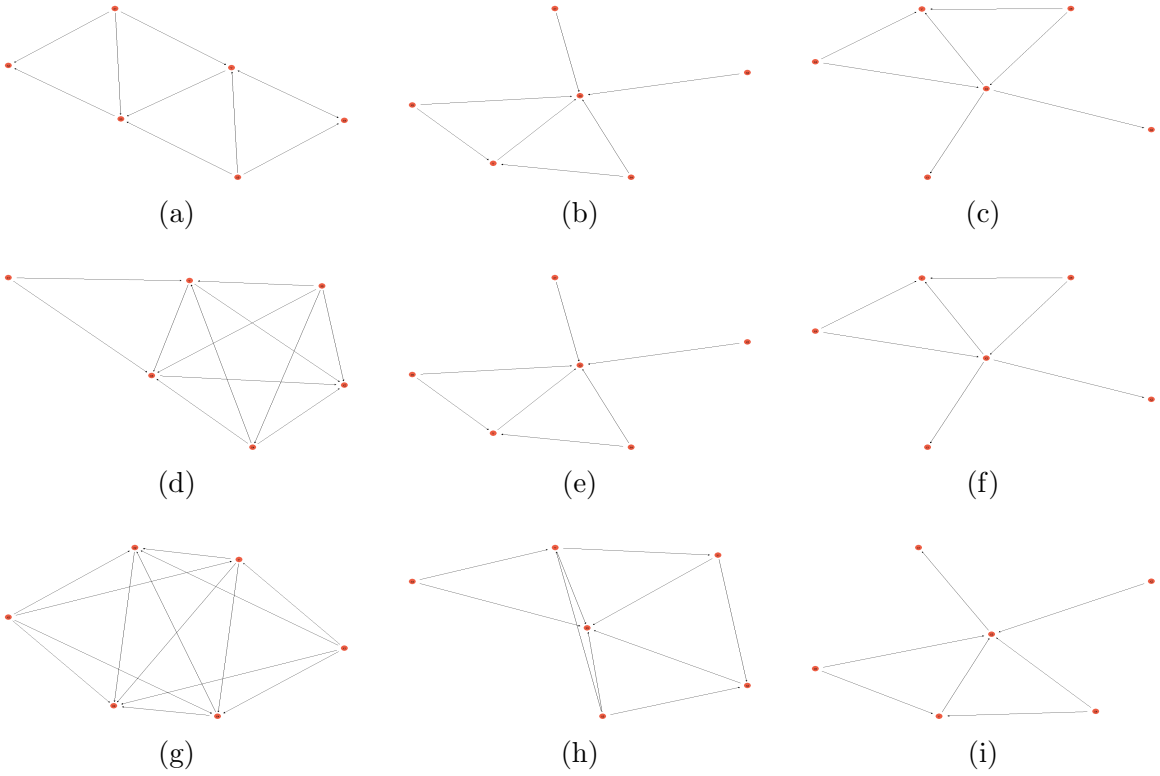

Figura 4: Estruturas obtidas por meio do método K2, para número de classes da variável resposta igual a 3.

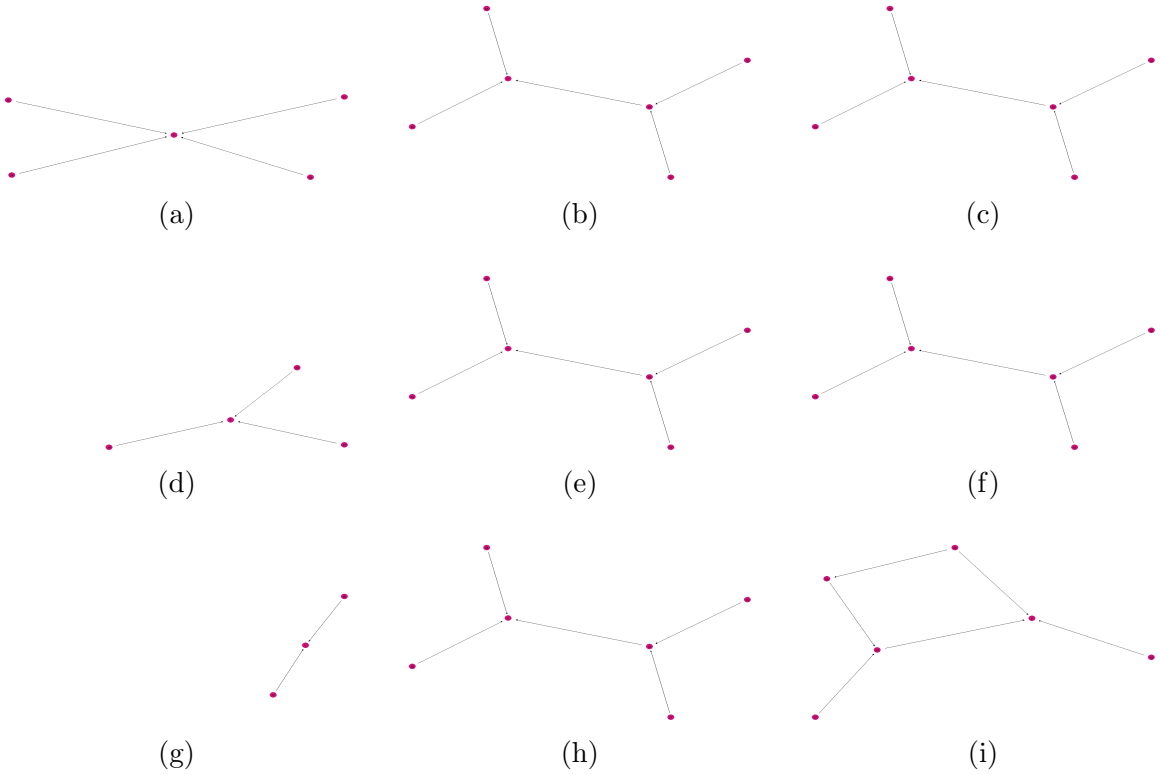

Figura 5: Estruturas obtidas por meio do método K2+PC, para número de classes da variável resposta igual a 3.

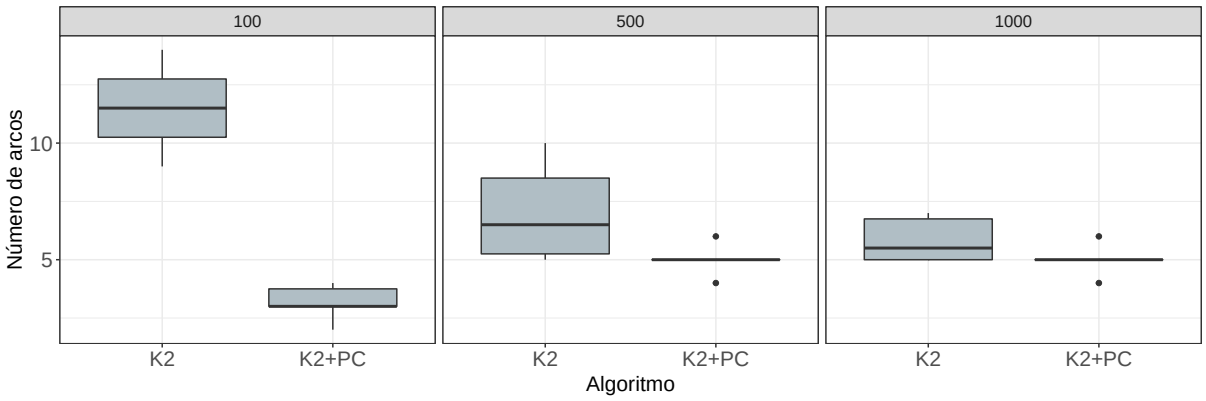

Figura 6: Distribuição do número geral de arcos atribuídos pelos algoritmos com base no tamanho amostral.

Aplicação para dados reais de agricultura

O conjunto de dados utilizado é composto de variáveis explicativas sobre o solo e o clima. Com relação ao solo, tem-se as variáveis tipo de solo, textura, pH, bem como quantidades de matéria orgânica presente, como fósforo, potássio, cálcio, magnésio, hidrogênio, alumínio, enxofre, sódio, soma de bases, capacidade de troca catiônica, corretivo de solo e saturação básica. Com relação ao clima, tem-se o mínimo, máximo e a média de temperatura, umidade do ar na região, média da radiação solar, velocidade do vento, e evapotranspiração. Além dessas variáveis, há a quantidade de fertilizantes utilizada, e seu tipo, fungicida, herbicida, inseticida, maturador; além da semeadura, sistema de plantação e variedade; bem como o tipo de colheita, irrigação e sistema de manejo. A variável de interesse é o *Percentual de Falha*, o qual se encontra em uma amplitude contínua, assim como a maioria das variáveis anteriores. Os dados são referentes a uma amostra aleatória de 2.748 talhões de cana-de-açúcar coletada entre Fevereiro de 2016 e Novembro de 2017, em diferentes regiões do Brasil. Uma visualização das variáveis consideradas para esta análise apresentadas na Figura 7. Além delas, a Variedade das plantas amostradas foi agrupada em 5 categorias (CT: 13,2%, CV: 10,7%, OT: 4,4%, RB8: 29,8% e RB9: 41,8%). Além disso, os dados continham alguns lotes com informações faltantes, os quais foram desconsiderados da análise. Os critérios de análise são os mesmos utilizados na Seção 4.

O problema em torno da base de dados se dá por encontrar as *causas* das falhas em lotes de plantas para fins industriais e comerciais, mais especificamente nos lotes onde a cana-de-açúcar é plantada. As Redes Bayesianas foram escolhidas para tal tarefa por conta da sugestão de causalidade dada pela orientação dos arcos no grafo acíclico e direcionado (PEARL, 2000). Então, uma vez que a estrutura de rede é estabelecida, se satisfizer certas condições discutidas anteriormente, na Seção , é possível definir os aspectos que influenciam/causam, direta ou indiretamente, uma característica de interesse, no caso, o *Percentual de Falha* em lotes de áreas plantadas.

Após a discretização das variáveis presentes na base de dados e descritas na Seção Método Híbrido, um método de seleção de variáveis é utilizado e baseado na métrica de Informação Mútua, dada pela expressão:

$$IM(X_i;Y) = \sum_{x_i \in X_i} \sum_{y \in Y} P(x_i, y) log \frac{P(x_i, y)}{P(x_i)P(y)},$$

a qual quantifica a informação compartilhada entre as duas variáveis X_i e Y. Os valores são calculados para cada uma das variáveis explicativas em relação a variável resposta, de modo que

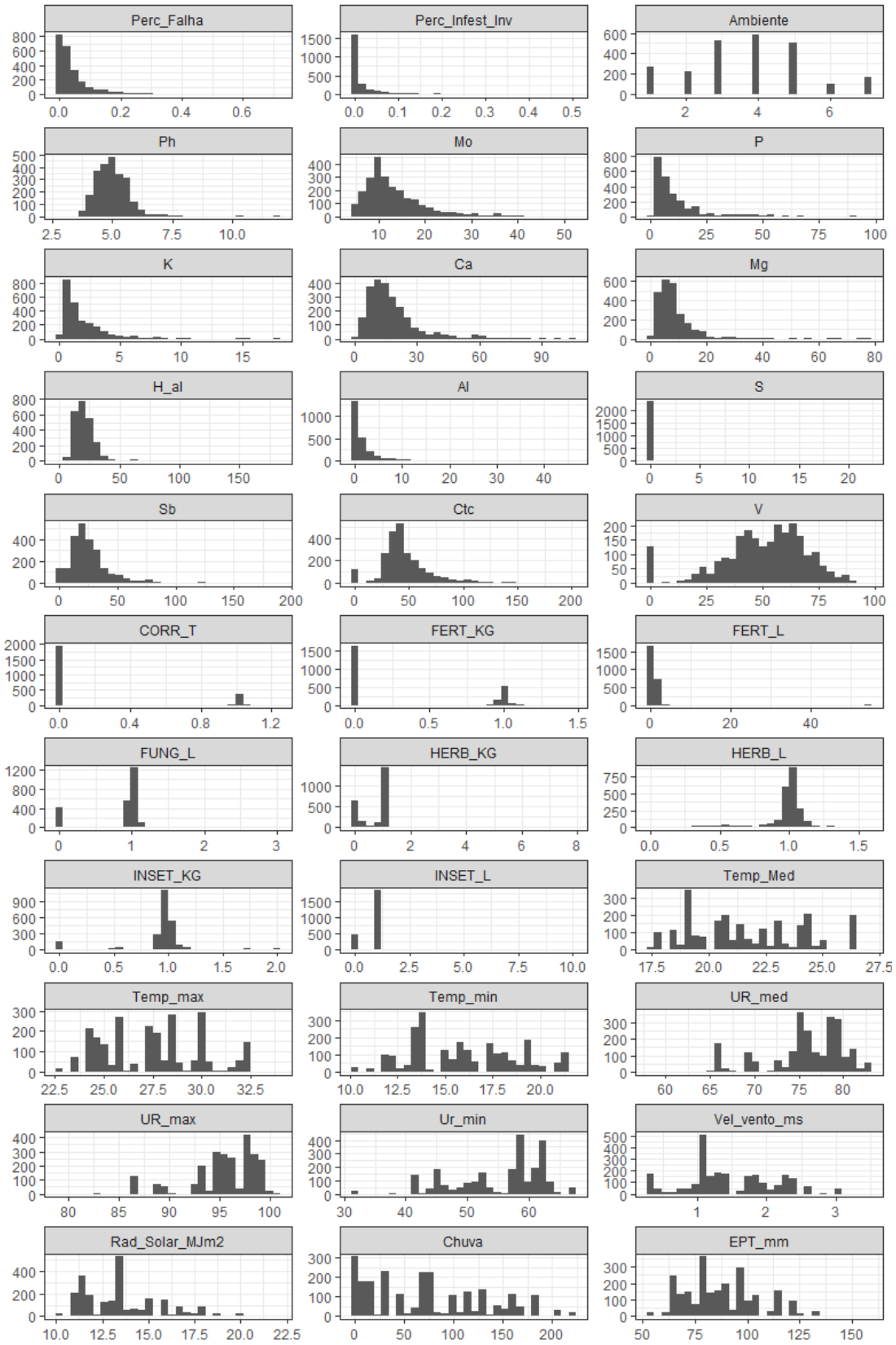

Figura 7: Distribuição das variáveis quantitativas do conjunto de dados agronômicos.

quanto maior for o seu valor, mais relevante é o atributo para a explicação/predição da variável de interesse.

Todo o processo descrito na Seção Método Híbrido foi aplicado na base de dados reais e a Tabela 2 contém os resultados das métricas de avaliação das redes, as quais estão apresentadas pela média ± desvio padrão - valores multiplicados por 100 - das 100 amostras 70/30 utilizadas para cálculo dos parâmetros e avaliação da predição, respectivamente. Por meio dessas métricas é possível comparar as seis configurações distintas testadas, as quais são definidas pela combinação da quantidade de classes que as variáveis explicativas e a variável resposta foram discretizadas.

Tabela 2: Tabela de resultados da estimação de rede por quantidade de categorias das variáveis discretizadas.

Classes		Algoritmo	Variáveis	Arcos	Acurácia	Coeficiente Correlacão de Mathews	Spherical Payoff
Y	X						
2	2	K2	13	71	-	-	-
		K2+PC	13	27	70,6 ± 1,8	70,7 ± 1,8	77,1 ± 1,2
	3	K2	6	30	-	-	-
		K2+PC	6	28	-	-	-
	4	K2	10	41	70,0 ± 1,7	70,0 ± 1,7	77,3 ± 1,0
		K2+PC	10	13	70,5 ± 2,0	70,5 ± 2,0	77,3 ± 1,0
3	2	K2	8	26	45,6 ± 1,9	59,5 ± 1,4	61,4 ± 0,9
		K2+PC	8	16	51,0 ± 1,8	63,4 ± 13,7	62,6 ± 1,1
	3	K2	9	36	-	-	-
		K2+PC	9	13	**53,2 ± 1,9**	**65,0 ± 1,4**	**63,2 ± 1,3**
	4	K2	9	32	-	-	-
		K2+PC	9	12	53,3 ± 2,0	65,0 ± 1,5	63,8 ± 1,5

Nesse resumo de resultados é possível notar a ausência de algumas medidas, isso se dá por conta da dificuldade do algoritmo para atribuir uma direção a aresta que liga duas variáveis. Esse não direcionamento dos arcos não permite que os parâmetros da Rede Bayesiana sejam calculados, uma vez que, ferem um dos aspectos que definem esse tipo de rede, que é o *DAG* - um grafo acíclico e direcionado.

Além disso, quando o algoritmo *PC* é aplicado recebendo como *whitelist* os pais resultantes da rede obtida pelo método *K2*, o número de arcos cai, em média, em 50%. Com a redução dessas conexões, a interpretação das relações entre as variáveis se torna mais viável. Adicionalmente, o acréscimo médio entre as medidas da qualidade preditiva dos métodos é de 3,3%, chegando a 5,6% de ganho na acurácia (quando essa comparação é possível); porém, em cerca de 67% dos modelos obtidos por meio da metodologia *K2* não é possível estimar os parâmetros devido ao não direcionamento completo dos arcos, diminuindo para 17% quando combinado ao método *K2+PC*.

Para ilustrar o ajuste gráfico, as Figuras 8 e 9 mostram as redes finais obtidas com cada uma das configurações e, respectivamente, correspondem ao método *K2* e *K2+PC*. As colunas das figuras correspondem a quantidade de categorias as quais as covariáveis discretizadas foram submetidas {2, 3, 4}, as linhas, por sua vez, a quantidade de categorias as quais a variável resposta foi discretizada {2,3}. Comparando a conectividade entre metodologias, os grafos obtidos pelo método *K2* independentemente da configuração, apresentam redes bastante conectadas, porém, quando o *PC* é aplicado e seus arcos reduzidos, a viabilidade de interpretação causal, ou meramente explicativa, das suas conexões está aliada ao aumento ou manutenção de sua capacidade preditiva.

Em uma análise de tempo computacional, pelo aumento no número de variáveis e, consequentemente, no número de parâmetros, houve um aumento no tempo investido nas iterações de amostras *bootstrap* para ajuste de um modelo de Redes Bayesianas. Na metodologia *K2* de 2,7 segundos até 3,7 minutos, em consequência disso, houve um aumento médio de 63,2%, bastante semelhante ao aumento obtido com os dados do estudo de simulação, o aumento absoluto variou de 1,29 a 48,1 segundos.

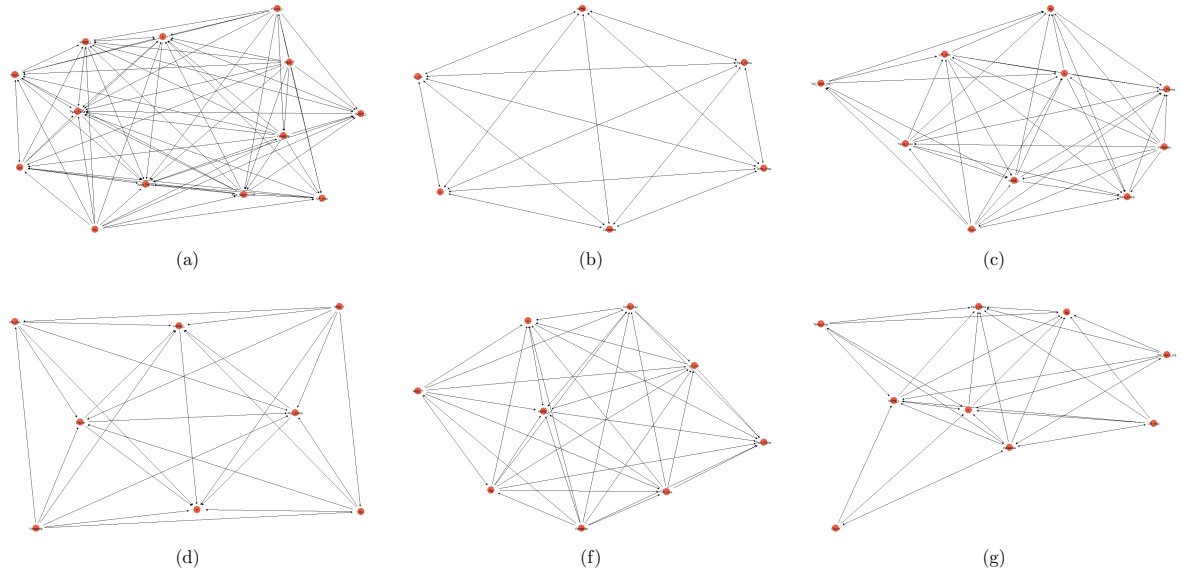

Figura 8: Estruturas obtidas por meio do método de estimação K2.

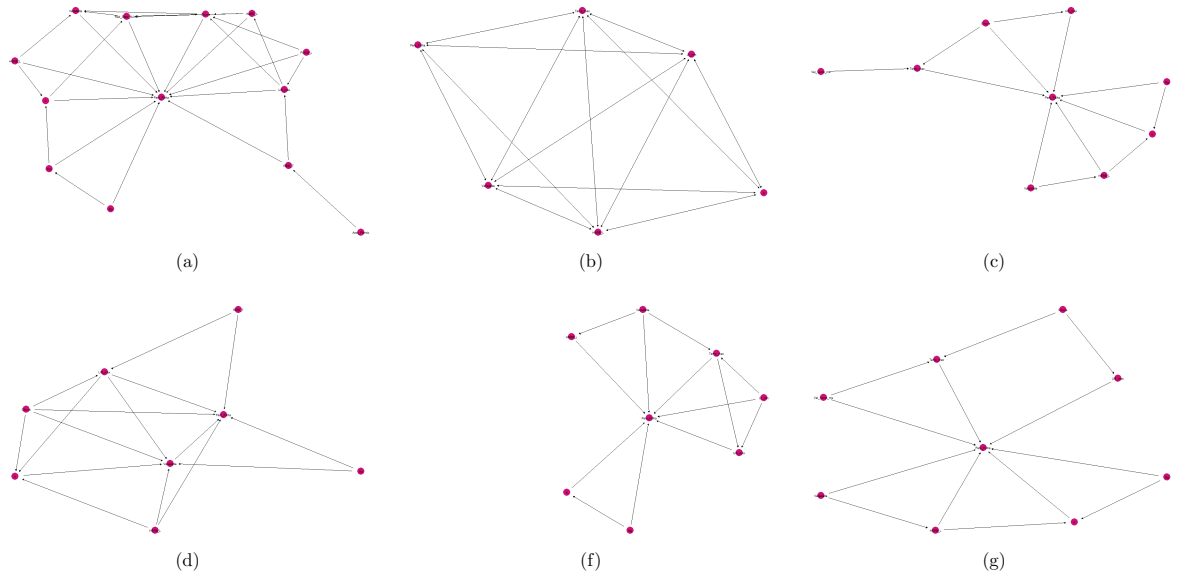

Figura 9: Estruturas obtidas por meio do método de estimação K2+PC.

Baseado nas estruturas acima, as redes que apresentaram melhor performance aliada a menor quantidade de conexões estão apresentadas na Figura 10 e Figura 11, respectivamente para 2 e 3 classes de discretização. Além de que seus arcos completamente direcionados fazem com que sejam Redes Bayesianas por definição. Contudo, ponderando ganho relativo em relação de uma estrutura aleatória de classificação, a Figura 11, com a resposta sendo discretizada em 3 categorias, obteve melhor desempenho do que a anterior e seu desempenho está destacado na Tabela 2.

As variáveis que estão presentes no grafo e foram submetidas ao procedimento de categorização, apresentam os seguintes pontos de corte: *Perc_Falha* $\{1,15; 4,17\}$; *K* $\{0,71; 1,26; 2,50\}$; *Mg* $\{4,08; 6,50; 9,35\}$; *HERB_L* $\{0,97; 1,01; 1,04\}$; *Vel_vento_ms* $\{1,08; 1,29; 1,86\}$; *Temp_max* $\{25,86; 27,73; 29,89\}$; *UR_max* $\{94,75; 95,91; 97,75\}$; *Chuva* $\{12,95; 68,90; 132,4\}$; sendo a variável *Variedade* originalmente de natureza categórica.

Esse grafo que é uma Rede Bayesiana, permite a análise dos aspectos que influenciam o

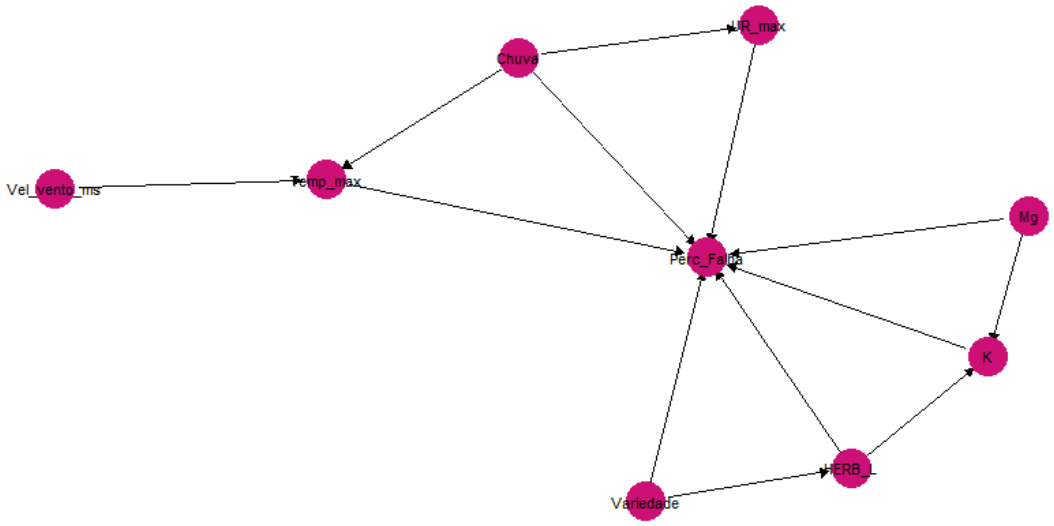

Figura 10: Estrutura final híbrida para a variável resposta discretizada em duas categorias.

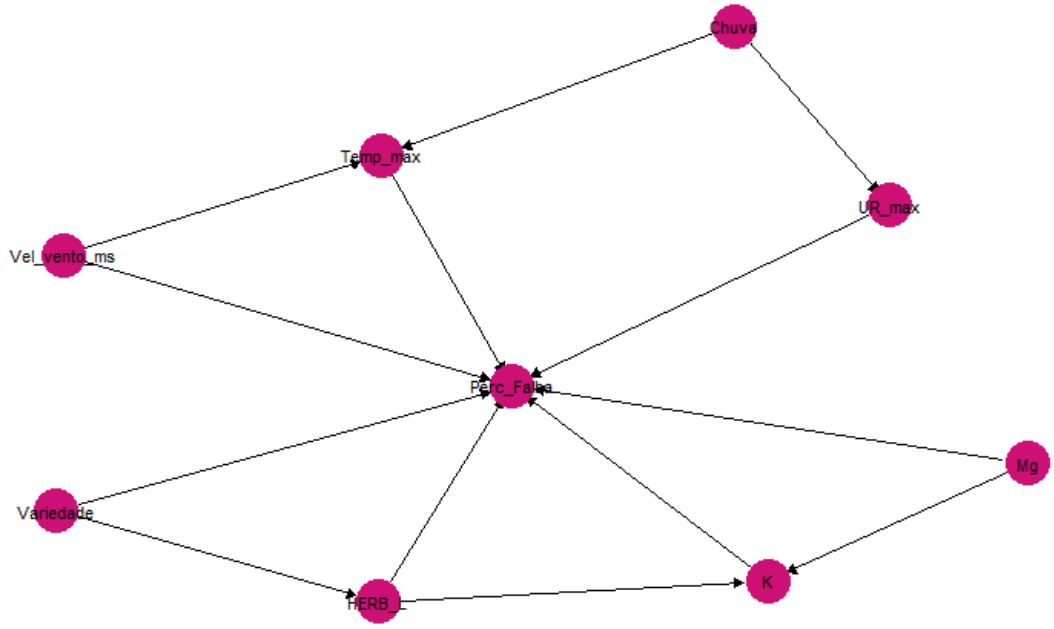

Figura 11: Estrutura final híbrida para a variável resposta discretizada em três categorias.

Percentual de Falhas, nesse cenário de variáveis. Ele possui sete pais, ou seja, existem sete variáveis que influenciam diretamente a variável resposta. Além de uma que não está ligada ao Percentual de Falhas diretamente, mas que se conecta a alguns de seus pais.

Em termos de causalidade, analisando as conexões, é possível perceber que a velocidade do vento (*Vel_vento_ms*), a temperatura máxima do local, em graus Celsius, (*Temp_max*), a umidade relativa máxima (*UR_max*), a quantidade de Magnésio (*Mg*), e Potássio (*K*) do solo, a quantidade de herbicida, em litros, (*HERB_L*) utilizado e a variedade causam diretamente a incerteza da variável resposta, bem como a chuva causa a variação na temperatura máxima e na

umidade relativa. Nota-se também que as variáveis climáticas estão mais agrupadas umas com as outras e que as variáveis relativas ao solo formam outro grupo.

Consideraçãoes finais

As Redes Bayesianas (RB), devido sua versatilidade, se adequam a diversos sistemas e podem auxiliar no entendimento de padrões e na tomada de decisão em diversos contextos. Contudo, nem sempre a estrutura de relação entre variáveis é conhecida e, portanto, deve ser estimada por conhecimento de especialistas ou por processos de aprendizado baseados em dados.

Nesse artigo apresentamos uma nova abordagem de uma metodologia híbrida de estimação de estruturas, denominada *scoring and restrict*, que busca por modelos parcimoniosos e poderosos preditivamente, por meio da união de dois algoritmos tradicionalmente eficientes. Nossa abordagem realiza a estimação de estrutura por meio da maximização da uma função objetivo, no caso a *K2*. E, a partir dessa estrutura inicial, as conexões diretas com a variável de interesse são supostas conhecidas. Passando, posteriormente, para a fase de testes de independência condicional com o *PC-stable*. Essa combinação resulta em redes equilibradas em termos de quantidade de conexões e, que ganham em desempenho preditivo. Através das análises apresentadas em dados artificiais, via procedimentos de simulação, essa combinação resulta em redes equilibradas em termos de quantidade de conexões e com aumento em capacidade preditiva. Assim, as redes moderadamente conectadas indicam conexões coerentes ao problema de identificação de aspectos que causam/influenciam as falhas em lotes de plantio de cana-de-açúcar, bem como apresentam uma arquitetura preditiva plausível de interpretação.

Considerando futuros trabalhos, outros métodos de estimação de estrutura baseados em diferentes métricas ou testes podem ser considerados, bem como diferente formas de combinação dos mesmos. Alternativas futuras a este estudo se baseiam em estudos de simulações mais amplos que envolvam outras configurações de heurísticas ou, até mesmo, comparações e estudo de influência de diferentes métodos de discretização.

Agradecimento

A pesquisa de Camila Ozelame foi financiada pela CAPES (Código de Financiamento 001). A pesquisa de Francisco Louzada é financiada pelo CNPq e FAPESP.

Referências

ABELLÁN, J.; GÓMEZ-OLMEDO, M.; MORAL, S. et al. Some variations on the PC Algorithm. *Probabilistic Graphical Models*. 2006, p.1–8.

ALVES, M.O.; FERREIRA, R.V.; ARAÚJO, G; BEZERRA, R. Otimizacao da Identificacao de Falhas de Plantio na Canda-de-Acucar com Usod de Geoprocessamento. *In*: X Congresso Brasileiro de AgroInformatica. 2015.

BERETTA, S.; CASTELLI, M.; GONÇALVES, I.; HENRIQUES, R.; RAMAZZOTTI, D. Learning the structure of Bayesian Networks: A quantitative assessment of the effect of different algorithmic schemes. *Complexity*, Hindawi, v.2018, 2018.

BIELZA, C.; LARRAÑAGA, P. Discrete Bayesian network classifiers: a survey, *ACM Computing Surveys (CSUR)*, v.47, n.1, p.5, 2014.

BOBBIO, A.; PORTINALE, L.; MINICHINO, M.; CIANCAMERLA, E. Improving the analysis of dependable systems by mapping fault trees into Bayesian networks. *Reliability*

Engineering & System Safety, Elsevier, v.71, n.3, 2001, p.249-260.

BUTTS, C. T. Network: a Package for Managing Relational Data in R. *Journal of Statistical Software*, v.24, n.2, 2008. Disponível em: http://www.jstatsoft.org/v24/i02/paper.

CAMPOS, L. M. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, v.7, p.2149-2187, Oct 2006.

COLOMBO, D.; MAATHUIS, M. H. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, JMLR org., v.15, n.1, p.3741-3782, 2014.

CONAB. Companhia Nacional de Abastecimento, safra 2018/2019. Levantamento Maio de 2018. *In*: Acompanhamento da safra brasileira de cana-de-açúcar, V1. ISSN: 2318-7921, 2018. Disponível em: https://www.conab.gov.br/info-agro/safras/.

CONAB. Companhia Nacional de Abastecimento, safra 2019/2020, Levantamento - Agosto de 2020. *In*: Acompanhamento da safra brasileira de cana-de-açúcar, V1. ISSN: 2318-7921, 2020. Disponível em: https://www.conab.gov.br/info-agro/safras/.

COOPER, G. F.; HERSKOVITS, E. A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, Springer, v.9, n.4, p.309-347. 1992.

DRURY, B.; VALVERDE-REBAZA, J.; MOURA, M.F.; LOPES, A.A. A survey of the applications of Bayesian networks in agriculture, *Engineering Applications of Artificial Intelligence*, Elsevier, v.65, p.29-42, 2017.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers, *Machine Learning*, v.29(2-3), p.131-163, 1997.

FRIEDMAN, N.; GOLDSZMIDT, M.; WYNER, A. Data analysis with Bayesian networks: A bootstrap approach. *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence*, p.196–201, 2013.

GÁMEZ, J. A.; MATEO, J. L.; PUERTA, J. M. Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood, *Data Mining and Knowledge Discovery*, Springer, v.22, n.1-2, p.106-148, 2011.

GASSE, M.; AUSSEM, A.; ELGHAZEL, H. An experimental comparison of hybrid algorithms for Bayesian network structure learning. *In*: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer. p.58-73, 2012.

GEIGER, D.; HECKERMAN, D. *Learning Gaussian networks*. Technical report, Microsoft Research, Advanced Technology Division, 1994.

GORODKIN, J. Comparing two K-category assignments by a K-category correlation coefficient. *Computational biology and chemistry*, Elsevier, v.28, n.5-6, p.367-374, 2004.

HARWOOD, J. L. *Managing risk in farming: concepts, research, and analysis*. US Department of Agriculture, ERS, 1999. 774p.

KOLLER, D.; FRIEDMAN, N. Probabilistic graphical models: principles and techniques. MIT press, 2009.

KORB, K. B.; NICHOLSON, A. E. *Bayesian artificial intelligence*, CRC press, 2010.

LERNER, B.; MALKA, R. Investigation of the K2 algorithm in learning Bayesian network classifiers, *Applied Artificial Intelligence*, Taylor & Francis, v.25, n.1, p.74-96, 2011.

LIU, Y.; MAN, H. Network vulnerability assessment using Bayesian networks. *In*: Data mining, intrusion detection, information assurance, and data networks security 2005. International Society for Optics and Photonics, v.5812, 2005, p.61-71.

LOUZADA, F.; ARA, A. Bagging k-dependence probabilistic networks: An alternative powerful fraud detection tool. *Expert Systems with Applications*, Elsevier, v.39, n.14, 2012, p.11583-11592.

MARCOT, B. G. Metrics for evaluating performance and uncertainty of Bayesian network models, *Ecological modelling*, Elsevier, v.230, p.50-62, 2012.

MEYER, P. E. *Infotheo: Information-Theoretic Measures*, R package version 1.2.0, 2014. Disponível em: https://CRAN.R-project.org/package=infotheo.

NADKARNI, S.; SHENOY, P.P. A Bayesian network approach to making inferences in causal maps. *European Journal of Operational Research*, Elsevier, v.128, n.3, 2001, p.479-498.

NAGARAJAN, R.; SCUTARI, M.; LÈBRE, S. *Bayesian networks in r*, Springer, v.122, p.125-127, 2013.

NEAPOLITAN, R. E. et al. *Learning bayesian networks*, Pearson Prentice Hall Upper Saddle River, NJ. v.38, 2004.

NIELSEN, T. D.; JENSEN, F. V. *Bayesian networks and decision graphs*, Springer Science & Business Media, 2009.

PEARL, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN: 0934613737.

PEARL, J. *Causality: models, reasoning and inference*, Springer, v.29, 2000.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. Disponível em: https://www.R-project.org/.

RASMUSSEN, S.; MADSEN, A. L.; LUND, M. Bayesian network as a modelling tool for risk management in agriculture. *IFRO Working Paper*. 2013.

RUSSELL, S.; NORVIG, P. *Artificial intelligence: a modern approach*, 2010.

RUZ, G. A.; ARAYA-DÍAZ, P. Predicting Facial Biotypes Using Continuous Bayesian Network Classifiers. *Complexity*, Hindawi, v.2018, 2018.

SCUTARI, M. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, v.35, n.3, p.1-22, 2010. DOI: 10.18637/jss.v035.i03.

SCUTARI, M.; DENIS, J. *Bayesian Networks with Examples in R*, Chapman and Hall, Boca Raton, 2014.

SCUTARI, M.; GRAAFLAND, C. E.; GUTIÉRREZ, J. M. Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, Elsevier, v.115, p.235-253, 2019.

SPIRTES, P.; GLYMOUR, C. N.; SCHEINES, R.; HECKERMAN, D. *Causation, prediction, and search*. MIT press, 2000.

STOLF, R. Metodologia de avaliação de falhas nas linhas de cana-de-açúcar. *Stab*, Piracicaba, v.4, n.6, 1986. p.22-36.

SU, C.; ANDREW, A.; KARAGAS, M.; BORSUK, M.E. Overview of Bayesian network approaches to model gene-environment interactions and cancer susceptibility. *6th International Congress on Environmental Modelling and Software*, 2012.

TSUBAKITANI, S.; EVANS, J. R. Optimizing tabu list size for the traveling salesman problem, *Computers & Operations Research*, Elsevier, v.25, n.2, p.91-97, 1998.

XUE, J.; GUI, D.; LEI, J.; SUN, H.; ZENG, F.; FENG, X. A hybrid Bayesian network approach for trade-offs between environmental flows and agricultural water using dynamic discretization. *Advances in Water Resources*, Elsevier, v.110, p.445-458, 2017.

YET, B.; CONSTANTINOU, A.; FENTON, N.; NEIL, M.; LUEDELING, E.; SHEPHERD, K. A Bayesian network framework for project cost, benefit and risk analysis with an agricultural development case study. *Expert Systems with Applications*, v.60, p.141-155, Elsevier. 2016.

ZHANG, X.; ZHAO, X.; HE, K.; LU, L.; CAO, Y.; LIU, J.; HAO, J.; LIU, Z.; CHEN, L. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, Oxford university press, v.28, n.1, p.98-104, 2012.