Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Mônica Pires Gravina de Oliveira

Identificação das Variáveis Relacionadas a Perda de Teor Alcoólico em uma Usina de Cana-de-Açúcar

Curitiba
2019

Mônica Pires Gravina de Oliveira

Identificação das Variáveis Relacionadas a Perda de Teor Alcoólico em uma Usina de Cana-de-Açúcar

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Walmes Marques Zeviani

Curitiba
2019

Identificação das Variáveis Relacionadas a Perda de Teor Alcoólico em uma Usina de Cana-de-Açúcar

Mônica Pires Gravina de Oliveira1,

1Departamento de Estatística, Universidade Federal do Paraná, Campus III - Centro Politécnico, Rua Evaristo F. F. da Costa, 418, Jardim das Americas, Curitiba - PR, Brasil*

Resumo

O processo de produção de etanol utilizando cana-de-açúcar como matéria-prima conta com diferentes etapas, tais como moagem da cana, tratamento do caldo, tratamento ácido do levedo, fermentação e destilação. Embora diferentes variáveis como contaminação bacteriana, floculação e viabilidade sejam conhecidas como prejudiciais à resposta de teor alcoólico, durante o processo fermentativo, são poucos os estudos relativos à associação dessas variáveis particularmente utilizando observações de uma usina de cana. As atuais ferramentas de controle de qualidade são incapazes de enfrentar a complexidade e as relações intrínsecas de variáveis deste processo, provendo valor limitado para tomada de decisão durante a safra. Assim sendo, o presente trabalho se propõe a utilizar regressão linear multivariada e árvores de regressão como abordagens para identificação das variáveis que influenciem negativamente a produção de teor alcoólico como primeiro passo para uma ferramenta que auxilie a tomada de decisão do setor.
Palavras-chave: usina de cana-de-açúcar, teor alcoólico, regressão linear multivariada, árvore de regressão

Abstract

The process of producing ethanol from sugarcane has different stages, from sugar cane milling, broth treatment, yeast acid treatment, fermentation and distillation. Although different variables such as bacterial contamination, flocculation and viability are known to be detrimental to the alcoholic response, during the fermentation process, studies related to the association of the different variables and this response using observations from a sugarcane mill are unknown in the literature. Current quality control tools are unable to cope with the complexity and intrinsic relationships of variables in this process, providing limited value for decision making during the harvest. Thus, the present work proposes to use multivariate linear regression and partition trees as approaches to identify the variables that negatively influence the production of alcohol content as a first step to a tool that assists decision making in the sector.
Keywords: sugarcane mill, alcohol content, multivariate regression model, partition tree

1. Introdução

A produção brasileira de cana-de-açúcar na safra 2016/2017 foi de 652 milhões de toneladas, resultando em 39 mil toneladas de açúcar e 27 mil m3 de etanol [1]. Com o aquecimento do setor pelas políticas do Renovabio, a safra de 2017-2018 foi majoritariamente voltada a produção de etanol. Tais políticas envolvem garantir o papel estratégico dos biocombustíveis apresentando três objetivos que incluem: cumprir os compromissos firmados no Acordo de Paris, incentivar a expansão dos biocombustíveis com foco na regularidade do abastecimento e garantir/ampliar a previsibilidade [2]. Uma usina de cana-de-açúcar típica é composta pelos processos de: recepção e preparo da cana e extração do caldo. O caldo é enviado para um sistema de tratamento, onde suas impurezas são removidas para fornecer material adequado para as etapas subsequentes. O caldo tratado segue para a etapa de cristalização do açúcar sendo o açúcar cristalizado centrifugado e lavado. A lavagem do açúcar gera um resíduo chamado melaço, que é utilizado na etapa de fermentação, onde ocorre adição de levedo. Após cada etapa de fermentação, o vinho obtido é levado para a destilação objetivando obtenção do etanol, enquanto o levedo sofre um processo de tratamento ácido para retornar ao sistema [3]. O processo de produção de etanol vem sendo amplamente estudado ao longo dos anos por meio de modelos fenomenológicos da etapa de fermentação [4]. Esses modelos são capazes de representar com boa precisão as variações das concentrações de células, substrato e etanol ao longo da fermentação, porém, não contemplam outras variáveis como contaminação bacteriana, floculação e viabilidade, que conhecidamente impactam nos resultados de teor alcoólico. As atuais ferramentas de controle de qualidade são incapazes de enfrentar a complexidade e as relações intrínsecas de variáveis deste processo, provendo valor limitado para tomada de decisão durante a safra. Portanto, faz-se necessário a utilização de outros modelos que possam contemplar a influência dessas variáveis no teor alcoólico fermentativo. Assim sendo, o presente projeto se propõe a investigar as variáveis associadas à produção de etanol de uma usina de cana-de-açúcar nas safras de 2016 e

*monica.gravina@gmail.com

2017, identificando as principais variáveis relacionadas à perda de teor alcoólico durante o processo fermentativo. A estratégia utilizada para estabelecer quais as variáveis de interesse foram regressão linear multivariada e árvore de regressão. Por não se tratar de um trabalho que objetiva gerar uma ferramenta preditiva, apenas descrever o presente conjunto de dados a avaliação independente do conjunto de dados não foi contemplada.

2. Métodos Multivariados

2.1. Regressão linear múltipla

A regressão linear múltipla é uma extensão da regressão linear simples, em que um conjunto de variáveis independentes são utilizadas para explicar a resposta. O modelo linear geral multivariado é dado por:

$$\underset{(nxm)}{\boldsymbol{Y}} = \underset{(nxk+1)}{\boldsymbol{X}}\ \underset{(k+1xm)}{\boldsymbol{\beta}} + \underset{(nxm)}{\boldsymbol{E}} \quad , (1)$$

Em que **Y** é a matriz de n casos de m variáveis resposta; **X** é um modelo matricial com colunas para k + 1 regressores, tipicamente incluindo uma coluna inicial de 1s para constante da regressão; β é uma matriz de coeficientes de regressão, uma coluna para cada variável explicativa; e **E** é a matriz de erros. A presunção de um modelo linear multivariado se preocupa com o comportamento dos erros: seja ε_i' a representação da iésima linha de **E**. Então $\varepsilon_i' \sim N_m(0,\Sigma)$, em que Σ é uma matriz não singular de erro- covariância, constante por meio dos casos; ε_i' e ε_i', são independentes para i ≠ i'; e **X** é fixado ou independente de
E. O estimador de máxima verossimilhança de **B** no modelo linear multivariado é equivalente a equação dos mínimos quadrados para cada resposta individualizada:

$$\hat{B} = (\mathbf{X'X})^{-1}\mathbf{X'}\mathbb{y}\ , \quad (2)$$

A condição de existência de $(\boldsymbol{X'X})^{-1}$ é que as colunas de $\boldsymbol{X}$ sejam linearmente independentes, ou seja, que nenhuma coluna de $\boldsymbol{X}$ seja combinação linear das demais. Na análise de variância em regressão linear múltipla, a variação total (corrigida pela média) é novamente decomposta em dois componentes (variação explicada pela regressão e variação residual), tal que:

$$SQTotal = SQReg + SQRes \quad , (3)$$

Usando notação matricial, as somas de quadrados ficam definidas por:

$$SQRes = \boldsymbol{y'y} - \hat{\beta}'\mathbf{X'y} \quad , (4)$$

$$SQReg = \hat{\beta}'\mathbb{X}\,\mathbb{y} - (\sum_{i=1}^{n} y_i)^2/n \quad , (5)$$

$$SQTotal = \dot{y}y - (\sum_{i=1}^{n} y_i)^2/n \quad , (6)$$

Podemos testar a significância do modelo ajustado com base no seguinte par de hipóteses:

$H0 : \beta 1 = \beta 2 = ... = \beta k = 0;$

$H1 : \beta j \neq 0$ para pelo menos em um j.

Sob a hipótese nula (não significância do modelo) a estatística F segue distribuição $F-$Snedecor, com $p - 1$ e $n - p$ graus de liberdade. Em que n é o tamanho da amostra e $p = k + 1$ o número de parâmetros do modelo. Assim, fixado um nível de significância α, a $H0$ deve ser rejeitada se o valor da estatística F for maior que o quantil $1 - \alpha$ da distribuição $Fp-1,n-p$. O coeficiente de determinação, que expressa a proporção da variabilidade original dos dados explicada pelo modelo de regressão, fica definido por:

$$R^2 = 1 - SQRes/SQTotal = SQReg/SQTotal \quad , (7)$$

Uma propriedade de $R2$ que o torna pouco apropriado para a comparação dos ajustes de diferentes modelos é que ele nunca decresce à medida que incluímos novas variáveis ao modelo. Como alternativa ao R^2 podemos considerar o R^2 ajustado, definido por:

$$R^2Aj = 1 - SQRes/(n - p)/SQTotal/(n - 1) \quad , (8)$$

Como $SQTotal/(n-1)$ é fixo, então R^2Aj somente aumentará se houver redução do quadrado médio de resíduos. Diferentemente de R^2, R^2Aj penaliza a inclusão de variáveis não importantes no modelo, permitindo comparar adequadamente modelos com diferentes complexidades (números de variáveis) [5].

2.2. Árvores de Decisão

Diferentes técnicas podem ser utilizadas para criar modelos que relacionam "entradas" (variáveis independentes) a uma "saída" (variável dependente). Entre as diferentes técnicas deve-se levar em consideração a capacidade de modelagem e a facilidade em interpretar o modelo criado. Árvores de decisão são estruturas do tipo árvore em que a bifurcação é representada por testes lógicos. No geral, os algoritmos de indução de árvores de decisão usam aproximações para quantificar a pureza, como o índice Gini e a entropia em tarefas de classificação, ou desvio padrão para tarefas de regressão. Assim, para um determinado conjunto de dados, o algoritmo varre todas as variáveis de entrada buscando maximizar a pureza da variável de saída em cada partição. O objetivo é realizar a partição da resposta entre grupos homogêneos, mas também fazer com que a árvore tenha um tamanho relativamente pequeno. Na representação de

árvore, o nó inicial representa todas as amostras do conjunto de dados utilizado para treinamento. Cada nó subsequente representa uma divisão do conjunto de dados com base em um teste lógico, até que seja atingido um critério de parada criando nós denominados "folhas". Para uma variável dependente numérica, a partição é definida por valores menores que e maiores que algum valor escolhido. Para cada "folha" da árvore, a predição do modelo é feita considerando a "saída" das amostras que foram alocadas na folha. Uma árvore de decisão é facilmente interpretável quando possui poucas "folhas" ou pouca profundidade (número de divisões até atingir uma "folha"). O tamanho da árvore é igual ao tamanho dos grupos finais. A partição é continuada até uma grande árvore ser construída, a qual é podada (*pruned*) até o tamanho desejado. Uma das formas recomendadas para a poda é por meio da escolha do número de nós que apresenta o menor erro de validação cruzada [6, 7, 8].

3. Ajuste dos modelos

3.1. Descrição do Conjunto de Dados

Os dados utilizados no estudo correspondem à uma série histórica da produção diária de etanol a partir de cana-de-açúcar em uma destilaria autônoma durante duas safras (Abril a Novembro de 2016 e 2017) perfazendo um total de 462 observações para o total das 50 variáveis registradas. As variáveis foram organizadas de acordo com as etapas do processo descritos a seguir: moagem, tratamento do caldo, tratamento ácido, fermentação e destilação (Tabela 1). A primeira etapa incluiu variáveis relacionadas a caracterização da matéria-prima e segunda incluiu a caracterização do mosto. Outras duas variáveis foram adicionadas sendo elas teor alcoólico calculado e delta. A primeira corresponde a uma relação, indicada pela Equação 9, entre o teor de açúcar redutor total no mosto (art mosto) em massa de mosto (m/m); o fator de conversão de açúcar em etanol (0,511 g/mL); a densidade de álcool (0,789 g/mL); o volume do mosto aqui descrito como volume de dorna subtraído do volume de cuba; o volume de dorna; e um fator de correção. Já a segunda corresponde à subtração de teor alcoólico calculado e o teor alcoólico real. A variável volume de cuba só foi observada para a safra de 2017, portanto o modelo gerado para o delta refletirá somente as informações obtidas em 2017.

Teor alcoólico calculado (v/v) = ART (0,511/0,789)*(V. dorna-V. cuba)/V.dorna + 5* , (9)

Etapa do Processo	**Variáveis de processo**
Moenda e caracterização da matéria-prima	moagem (ton/h), art entrada (kg/ton), atr mp, impureza mineral mp, impureza vegetal mp, ph caldo pcts, dextrana mp,
Caracterização do mosto	art mosto, sólidos mosto, ph mosto, temperatura mosto
Tratamento Ácido	teor creme, ph fermento, tempo de tratamento ácido (h), acidez fermento, ácido sulfúrico (litros), ácido sulfúrico (kg/m³ et), vazão mosto (m³/h) dispersante (litros), dióxido de cloro (litros), dispersante (g/ m³ et), dióxido de cloro (g/m³ et)
Fermentação	teor alcoólico, arrt vinho centrifugado, glicerol v.c. volume de cuba volume de dorna, bast/ml final, floculação ferm, viabilidade final, temperatura máxima vinho bruto, tempo de alimentação (h), tempo espera centrifugação (h), tempo de fermentação (h), antiespumante (litros), antiespumante (g/m³ et), % etanol recuperado em relação etanol total produzido, rendimento fermentativo, rendimento fermentativo art
Destilação	etanol 100%, etanol total, condutividade hidratado, condutividade anidro, rendimento destilação rendimento geral da

Tabela 1. Apresentação das variáveis disponíveis no conjunto de dados.

3.2 Tratamento Preliminar dos Dados

A etapa de tratamento preliminar dos dados incluiu a padronização dos nomes das variáveis além de remoção de

caracteres especiais presentes no conjunto de dados. Nesta etapa e nas demais foi utilizada linguagem R [9]. Para o processamento dos dados foram utilizadas as bibliotecas *stringi* [10] e *stringr* [11].

3.3 Análise Exploratória e Remoção de Outliers

Os dados foram distribuídos em histogramas para verificação de possíveis *outliers* (Figura 1). As variáveis relacionadas à concentração de açúcar (ART de entrada, atr mp e ART do mosto) e ao pH (pH do caldo PCTS e pH do mosto) apresentaram valores extremos menores que ou equivalentes a 0, tendo sido considerados *outliers*, valores impossíveis de se observar em uma usina de cana. Os dados também foram divididos em quartis, de modo que foram verificados os valores mínimos de cada variável e outros *outliers* foram identificados sendo estes substituídos por *NAs* de acordo com regra da Tabela 2. Pontos mínimos e máximos extremos não foram removidos do conjunto de dados, tendo em vista o interesse em descrevê-los por meio dos modelos aqui investigados. O pacote *dplyr* foi utilizado nesta etapa [12].

Variáveis Filtradas	Filtros
Impureza Mineral da Matéria- Prima; Impureza Vegetal da Matéria-Prima; Dextrana da Matéria-Prima; pH do Caldo PCTS; ATR Matéria-Prima; ART Mosto; pH do Mosto; Acidez Final; Teor do Creme; Glicerol do Vinho Centrifugado; ART de Entrada (kg/ton); Temperatura do Mosto; Ácido Sulfúrico (L); Rendimento Fermentativo; Rendimento Fermentativo ART; Rendimento Destilação; Rendimento Geral da Destilaria; Condutividade Etanol Anidro; Condutividade Etanol Hidratado;	Menor ou Igual a 0
Bastonete/mL	Menor ou Igual a 10
Temperatura Máxima Vinho Bruto	Menor ou Igual a 15
Dióxido de Cloro (L)	Menor ou Igual a 1

Tabela 2. Filtros usados para remoção de potenciais *outliers*. Os outliers foram substituídos por *NAs*.

Após a remoção dos outliers, os dados foram distribuídos em séries temporais (Figura 2), bem como em gráficos que apresentavam as variáveis de interesse (teor alcoólico real e delta) frente às outras variáveis processuais para verificação inicial de relações diretas ou indiretas entre as variáveis (Figura 3). A análise preliminar mostrou principalmente que variáveis relacionadas à concentração de açúcar (ART entrada, atr mp e ART mosto) seguem um padrão de distribuição similar ao do teor alcoólico ao longo das duas safras. Isto pode ser justificado pelo fato do açúcar redutor apresentar uma relação estequiométrica com a concentração de etanol, sendo diretamente representado por um fator de conversão de 0,511 [13]. As variáveis resposta (teor alcoólico e delta) foram distribuídas em um gráfico de pontos frente às diferentes variáveis de processo. Para a variável teor alcoólico fica claro que, além das previamente citadas, as variáveis glicerol, impureza mineral, impureza vegetal, moagem, pH do fermento e tempo de alimentação apresentaram aparente correlação linear com a variável resposta (Figura 4). O gráfico da diferença entre teor alcoólico calculado e real (delta) ao longo das safras (Figura 5), indica que as maiores variações na resposta podem ser verificadas no início e no fim da safra, possivelmente sendo relacionadas a épocas de maior precipitação (mm), que se dão nos período de março a maio e setembro a novembro. Além disso, o início da safra é caracterizado por falta de estabilidade operacional incluindo paradas no processo e alta variabilidade na resposta de teor alcoólico.

3.3.1 Matriz de Correlação Linear

Foram realizados testes de correlação linear entre as variáveis de interesse (art mosto, teor alcoólico e delta) e suas potenciais variáveis explicativas por meio da biblioteca *corrplot* [14]. O objetivo principal foi auxiliar na seleção das variáveis que seriam utilizadas na geração dos modelos multivariados. As variáveis que apresentaram correlação linear significativa ($\alpha = 0,25$) foram incluídas na matriz de correlação, cuja representação da intensidade da correlação é dada por uma escala de cores onde azul indica uma correlação positiva e vermelho uma correlação negativa. A significância estabelecida foi acima da faixa comum (0,05- 0,1) para que as variáveis explicativas não fossem eliminadas no processo automático de seleção. As principais correlações observadas foram as das variáveis ART mosto, teor alcoólico e delta (Figura 6). Com base no gráfico de correlação pode-se verificar que todas as variáveis explicativas (moenda/caracterização da matéria-prima e caracterização do mosto) da variável ART do mosto apresentaram correlação linear significativa, embora somente moagem (ton/h), açúcar total da matéria-prima (atr mp) e impurezas (mineral e vegetal) tenham apresentado correlação moderada ($r = 0,5$ a $0,6$). Assim como para art do mosto, as potenciais variáveis explicativas do teor alcoólico (moenda/caracterização da matéria-prima, caracterização do mosto, tratamento ácido e fermentação) também apresentaram correlação linearmente significativa, excetuando-se dióxido de cloro e tempo de fermentação. Dentre as primeiras, as variáveis que apresentaram correlação linear moderada ($r = 0,5$ a $0,6$) foram acidez final, moagem (ton/h), atr mp, art do mosto, impureza mineral e vegetal, pH do fermento e vazão do mosto. A correlação foi negativa para as impurezas e pH do fermento, sendo positiva para as variáveis restantes. Por fim, observou-se a variável delta, para a qual verificaram-se as seguintes correlações significativas: ácido sulfúrico, acidez final, glicerol do vinho centrifugado, vazão do mosto, antiespumante, dispersante, tempo de tratamento ácido, açúcar residual total do vinho, viabilidade,

temperatura máxima do vinho bruto, dióxido de cloro, teor de levedo, tempo de alimentação e de fermentação, bastonete/mL, impureza mineral e vegetal, floculação, pH

3.4 Modelos de Regressão Linear Multivariada – ART do mosto, Teor alcoólico observado e Delta

Inicialmente, as variáveis foram padronizadas de acordo com método min-max (Equação 2). À variável bastonete/mL foi também aplicada uma transformação logarítmica.

Valor Normalizado (i) = (Valor Observado – Valor Min Observado)/(Valor Max Observado – Valor Min Observado) , (*10*)

Utilizando as variáveis padronizadas, foram gerados três modelos para o açúcar redutor total (ART) no mosto, para o teor alcoólico observado e para o delta (teor alcoólico calculado – real). As variáveis significativas para o açúcar no mosto compunham o modelo de teor alcoólico em uma estratégia sequencial. O procedimento de seleção de variáveis foi realizado de forma iterativa, adicionando e removendo variáveis, com base em um critério de seleção por meio do método *stepwise*. Os critérios de seleção de variáveis utilizados foram o índice de akaike (AIC), bem como o p-valor ($\alpha = 0,05$), sendo utilizada a biblioteca *olsrr* [15]. Neste caso o nível de significância utilizado foi menor do que o utilizado no método de correlação linear, pois neste último objetivou-se selecionar variáveis de maneira não tão restritiva, neste caso pretende-se identificar por método automatizado as variáveis mais relevantes. O método VIF (*Variance Inflation Fator*) também foi utilizado para testar a multicolinearidade das variáveis por meio do pacote *car* [16]. Em caso de multicolinearidade severa, eliminaram-se as variáveis explicativas altamente correlacionadas ($VIF \geq 10$). A linearidade das respostas frente a cada variável explicativa foi verificada por meio de distribuições marginais de resíduos de *Pearson*.

3.4.1 Modelo de ART do Mosto

Para explicar a resposta de % (m/m) de açúcar total presente no mosto, foi gerado um modelo com base nas variáveis presentes nas etapas de moagem e caracterização do mosto sendo estas selecionadas utilizando o método *stepwise* (moagem (ton/h), pH do caldo pcts, sólidos do mosto, dextrana da mp, atr da mp, impureza mineral da mp, ph do mosto, temperatura do mosto). O modelo final apresentava Moagem de cana (ton/h), açúcar total da matéria-prima, impureza mineral da matéria-prima, sólidos do mosto e temperatura do mosto (Figura 7) como variáveis explicativas da resposta ART do mosto. A significância estatística do modelo testada por meio do teste F indicou que o modelo é significativo ($\alpha = 0,05$). O ajuste do modelo aos dados também foi expresso pelo coeficiente de determinação ajustado, R^2, o qual foi de 0,716. Este valor indica que 71,6% das variações na resposta açúcar redutor total (art) do mosto são explicadas pelo modelo. Em meio às variáveis explicativas, o efeito positivo da variável moagem pode ser explicado pois quanto mais rápido a cana é limpa, lavada e processada na moenda, menos tempo ela fica na presença de contaminantes que podem resultar na redução da concentração de açúcares no caldo da cana. Os principais micro-organismos que causam a degradação da cana, uma vez cortada, são as bactérias produtoras de ácido láctico, sendo estimadas perdas diárias de sacarose da ordem de 4,75% [17]. De forma oposta, o aumento do teor de impurezas minerais interfere na qualidade do produto além de causar desgastes em equipamentos. O aumento da terra na cana contribui para a diminuição do açúcar principalmente durante a lavagem da cana e no processamento da torta de filtro. As impurezas minerais estão diretamente ligadas à perda de sacarose durante o tratamento do caldo e favorecem o aumento da produção de torta de filtro (resíduo) por tonelada de cana processada. Além disso, a terra carreia microrganismos que prejudicam a fermentação tais como leveduras selvagens as quais causam sérios prejuízos relacionados a sua baixa eficiência fermentativa/geração de floculação nas dornas culminando no aumento do gasto de insumos [18]. Outra variável significativa que apresentou estimador negativo foi a temperatura do mosto. O controle de temperatura na usina envolve gastos energéticos e de água. No entanto o aumento da temperatura do mosto pode impactar negativamente a concentração de açúcar na cana, conforme indicado no modelo. Esse efeito pode estar associado a contaminação microbiana do mosto dado que temperaturas em torno de ~35-40ºC favorecem proliferação de contaminantes como bactérias lácticas [19], presentes no processo, culminando em redução do açúcar no caldo. A análise de resíduos não indicou padrões sistemáticos, confirmando a relação linear e adequação do modelo à resposta de art do mosto (Figura 10).

3.4.2 Modelo de Teor Alcoólico

O modelo do teor alcoólico presente no vinho bruto foi gerado com base nas variáveis originadas do modelo para o açúcar redutor do mosto, e aquelas presentes nas etapas de fermentação e de tratamento da levedura. Apesar de apresentar correlação linear moderada e significativa, vazão do mosto não será considerada como variável explicativa do modelo de teor alcoólico tendo em vista que só foi registrada no período de 2017. As vinte e duas variáveis investigadas foram moagem (ton/h), sólidos do mosto, açúcar total da matéria-prima, impureza mineral da matéria- prima, temperatura do mosto, acidez final, teor do creme, bastonete/ mL, viabilidade final, floculação do fermento, glicerol no vinho centrifugado, pH do fermento, ácido sulfúrico (L), teor de levedo, temperatura máxima do vinho bruto, tempo de tratamento ácido (h), tempo de alimentação (h), tempo de espera para centrifugação (h), açúcar residual do vinho centrifugado, volume de dorna, dióxido de cloro (L), antiespumante (L) e dispersante (L). O método *stepwise* foi utilizado para seleção das variáveis, sendo as seguintes

selecionadas: açúcar total da matéria-prima, impureza mineral da matéria-prima, acidez final, floculação, glicerol, pH do fermento, teor de levedo, temperatura máxima do vinho bruto e tempo de tratamento ácido. A significância estatística do modelo testada por meio do teste F indicou que o modelo é significativo ($\alpha = 0,05$). O ajuste do modelo também foi expresso pelo coeficiente de determinação ajustado, R^2_a, o qual foi de 0,717. Este valor indica que 71,7% das variações na resposta açúcar redutor total (art) do mosto são explicadas pelas variáveis componentes do modelo escolhido (Figura 8). Dentre as variáveis significativas, acidez final, floculação, glicerol, pH do fermento, teor de levedo, temperatura máxima do vinho bruto e tempo de tratamento ácido devem ser destacadas, visto que as relações entre impureza mineral e açúcar da matéria prima frente ao seus impactos no processo já foram previamente explicadas nos tópicos 3.3 e 3.4.1. O estudo de como cada um desses principais fatores influencia o resultado de teor alcoólico é importante para o entendimento dos mecanismos relacionados à eficiência da levedura em produzir etanol. Segundo Atala [20], os principais fatores que limitam a produtividade das leveduras, responsáveis pela fermentação alcoólica deste processo são: a) Temperatura; b) Acidez (pH); c) Nutrientes; d) Concentração de Etanol; e) Agentes tóxicos;

f) Pressão osmótica e; g) Contaminação microbiana [21]. Conforme observado, acidez final e glicerol apresentam estimadores positivos, sendo associados ao aumento da resposta de etanol dado o seu próprio incremento. Os mecanismos que justificam estas respostas estão relacionados ao estresse celular por vias diferentes. Enquanto o primeiro é associado a estresse pela presença de ácidos orgânicos que são produzidos por bactérias contaminantes, o segundo é associado ao estresse osmótico. A presença de glicerol é associada a dois fatores: ao crescimento celular e estresse osmótico. Em meio hipertônico, por exemplo em alta concentração de açúcares a levedura se encontrará em condição de estresse osmótico, e desenvolverá uma rápida resposta molecular para reparar danos e proteger suas estruturas celulares dos efeitos de estresse. Nesta condição, a célula rapidamente começa a perder água e sintetiza trealose e glicerol para se proteger a célula da desidratação e proteger as estruturas celulares do efeito dos efeitos da condição de estresse [22]. Como estes processos são catabólicos, ou seja, exigem gasto energético para ocorrerem, existe concomitante produção de etanol. Da mesma maneira a variável acidez final está associada a níveis de ácidos orgânicos ao fim da fermentação tais como acético, láctico, pirúvico e succínico [23]. Além de serem produzidos pela própria levedura (p.e. ácido acético) outro fator ligado a níveis elevados de ácidos orgânicos ao fim da fermentação remete à contaminação por bactérias ácido- lácticas, produtoras de ácido láctico e ácido acético. Um dos mecanismos de estresse por ácidos orgânicos remete a capacidade destes atravessarem a membrana celular da levedura, chegando ao citosol da célula, reduzindo o pH, e podendo aumentar a atividade de H^+-ATPase na membrana celular ("bomba de prótons"), promulgando a geração de ATP (energia) que mais uma vez culmina na produção de etanol [24]. A temperatura máxima do vinho bruto também apresentou um estimador positivo para resposta de etanol podendo ser associada ao fato das leveduras possuírem temperatura ótima para incrementar a taxa de conversão do açúcar em etanol, no entanto outro mecanismo de estresse pode estar associado a essa resposta, uma vez que o choque térmico pode causar um pico inicial de atividade metabólica e resultar em maiores concentrações de etanol [25]. É válido ressaltar que embora estresse promova geração de etanol, a levedura não consegue permanecer por ciclos fermentativos sequenciais com níveis de estresse elevados, muitas vezes perdendo viabilidade resultando em morte celular. Os tempos de alimentação, tratamento ácido e espera para centrifugação correspondem ao tempo total de fermentação. Embora hoje muitas usinas adotem um diagrama de ocupação que indica a sequência de ocupação das dornas de fermentação e das cubas de tratamento ácido, com tempos associados a cada uma das etapas de tratamento ácido, alimentação de mosto, fermentação e espera para a centrifugação, estes tempos podem variar conforme o funcionamento da usina. Por exemplo, uma fermentação pode demorar mais tempo caso a centrífuga esteja parada ou sobrecarregada. No entanto, como o tratamento ácido não é seletivo para o seu alvo, quando a duração deste é excessiva, a levedura que é capaz de manter sua homeostase de forma quase independente dos valores de pH do tratamento ácido (2,0 a 3.2) podem sofrer danos e perda de viabilidade neste processo [22]. Isto culmina na necessidade de manutenção celular e decorrente gasto de carbono em massa celular, e não etanol. De forma oposta vemos o incremento do pH do fermento como negativamente influente para a resposta de produção de etanol. Isto pode estar associado a elevados índices de contaminação presentes na fermentação pois quando o pH não alcança o alvo de 2,0 a 3,2, o controle bacteriano e de leveduras selvagens não é efetuado de forma eficiente [26]. O teor de levedo está relacionado à proporção de levedura no vinho bruto. A causa pode estar associada à maior proporção de levedo tratado/mosto no início da fermentação e/ou ainda à maior crescimento da levedura durante a fermentação. Entretanto qualquer uma das alternativas culmina em mais açúcares convertidos em crescimento ou manutenção celular e consequente redução no teor alcoólico. Embora a análise de resíduos tenha indicado comportamento similar ao quadrático para algumas variáveis, verificou-se que estes padrões eram provocados por variáveis extremas, não culminando em necessidade de exploração de diferentes ordens das variáveis aqui presentes (Figura 11).

3.4.3 Modelo do Teor Alcoólico Real – Calculado (Delta)

Para explicar a resposta de teor alcoólico real – calculado (delta) foi gerado um modelo utilizando as variáveis presentes nas etapas de moagem, caracterização do mosto, fermentação e tratamento ácido selecionadas com base na matriz de correlação. As variáveis incluídas na etapa de seleção foram: ph do fermento, acidez final, ácido sulfúrico (L), glicerol do vinho centrifugado, antiespumante (L), dispersante (L), açúcar redutor do vinho centrifugado, viabilidade final, temperatura máxima do vinho bruto, tempo de alimentação (h), tempo de tratamento ácido (h), impureza mineral e vegetal da matéria-prima, teor levedo, dióxido de cloro (L), bastonete/ mL e floculação. O modelo final apresentava somente as variáveis acidez final e temperatura máxima do vinho bruto como variáveis

explicativas da resposta. A significância estatística do modelo testada por meio do teste F indicou que o modelo é significativo ($\alpha = 0,05$), no entanto o coeficiente de determinação ajustado, R^2_{adj} foi de 0,15. Este valor indicou que 15% das variações na resposta são explicadas pelo modelo (Figura 9). A baixa capacidade explicativa do modelo pode indicar que as variáveis utilizadas para calcular o etanol teórico (teor alcoólico calculado) não foram suficientes para representar esta variável. Por exemplo a variável densidade do mosto era necessária para um cálculo mais preciso, no entanto ela mesma não se encontrava nas observações cedidas pela usina. Também deve-se lembrar que art do mosto, volume de cuba e dorna são provenientes de medições de processo, as quais podem apresentar falhas relativas à mensuração. Por fim, o número de registros de volume de cuba restringia-se à safra de 2017, culminando em menor número de observações. No entanto duas variáveis foram significativas para este modelo, sendo elas acidez final e temperatura máxima do vinho centrifugado. Ambas foram também significativas para o modelo de teor alcoólico real, indicando que o delta reduz no incremento destas duas variáveis.

3.5 Árvore de Decisão

Análises exploratórias e modelos estatísticos comumente falham em encontrar padrões significativos em dados complexos, pois relações entre variáveis podem não ser lineares e envolver interações de alta ordem [7]. Tendo por objetivo utilizar outra metodologia que considere interações entre as diferentes variáveis com efeito nas respostas, a árvore de regressão foi utilizada. A poda foi efetuada por meio da escolha do número de nós que apresentasse o menor erro de validação cruzada disponibilizado pela biblioteca *rpart* que foi utilizada para este fim [27]. A biblioteca *rpart.plot* [28] também foi utilizada.

3.5.1 Árvore de Decisão do ART do mosto

Na figura 13 consta a árvore obtida com o conjunto de variáveis utilizadas no tópico 3.4.1 (modelo de art do mosto). A árvore gerada é consistente com o modelo linear multivariado gerado (Figura 7), em que se verifica a presença de moagem, impureza mineral, atr mp e sólidos do mosto como variáveis explicativas do % art do mosto. O erro relativo de validação cruzada é de 29,43%, percebendo- se que a árvore gerada é suficientemente representativa de divisões naturais do conjunto de dados. Na raiz da árvore temos a variável moagem como principal determinante da concentração de açúcar no mosto. Vê-se que valores de moagem menores do que 12000 ton de cana/h resultaram em um percentual de açúcar no mosto $\leq$ 9%. No ramo ao lado esquerdo da árvore, nota-se que a variável impureza mineral foi associada a menores percentuais de açúcar no mosto quando sua concentração na matéria-prima é $\geq$ 25%. Já nos ramos do lado direito, tem-se as variáveis açúcar na matéria-prima e sólidos do mosto como definidoras da concentração de açúcar no mosto, enquanto nota-se que a variação na concentração de açúcar na matéria prima < 141 kg/ton gerou uma quantidade de açúcar no mosto desde 16 a 17%, ao passo que maiores presenças de açúcar com teores de sólidos inferiores a 0,15% geraram 17% teor de açúcar no mosto.

3.5.2 Árvore de Decisão do Teor Alcoólico

Apesar de apresentar algumas variáveis extras para descrição do teor alcoólico, a árvore gerada (Figura 14) foi consistente com o modelo linear multivariado gerado (Figura 8). Neste caso, o erro relativo de validação cruzada foi de 22,17%, entendendo-se que a árvore gerada foi suficientemente representativa de divisões naturais do conjunto de dados. Na árvore verifica-se a presença das variáveis acidez final, arrt do vinho centrifugado, atr mp, impureza mineral, atr mp, moagem ton/h, pH do fermento e tempo de tratamento ácido. Na raiz da árvore temos a variável acidez final, indicando que valores de ácido >1,2% resultaram em maiores teores alcoólicos. No ramo ao lado esquerdo da árvore, nota-se a variável arrt do vinho centrifugado seguida de uma nova partição em que se vê a variável impureza mineral. Embora estas duas variáveis não sejam aparentemente relacionadas, a medição do açúcar residual do vinho (Brix) por meio de refratômetros não é restrita a medição de açúcares, mas também de sujidades presentes no mosto que por sua vez podem ser provenientes das impurezas minerais. Nota-se que na presença de impurezas minerais >_19%, o valor de teor alcoólico assumiu 4%. Os ramos a direita apresentam um nó inicial em que atrmp < 141 particiona-se e podem ser notados dois padrões de homogeneidade entre as variáveis. O ramo à esquerda de atrmp, apresenta partições contendo as variáveis moagem e impurezas minerais que são associadas geralmente a paradas da usina devido a chuvas, sendo de ocorrência comum ao início ou fim da safra. O ramo à direita apresenta variáveis associadas ao tratamento ácido incluindo o tempo do tratamento e pH do levedo tratado. Vê-se que quando o tempo de tratamento ácido era $\leq$ 2.3 (h), então o teor alcoólico assumia valores de 10% caso a moagem fosse >_16000 (ton/h). Caso o valor do tempo de tratamento ácido fosse $\geq$ 2,3 (h), então uma nova partição era gerada com o pH do fermento, verificando-se que um pH acima de 2,8 resultava em etanol ao redor de 7,9%.

3.5.3 Árvore de Decisão do Delta

A árvore representativa da resposta de delta apresentou apenas duas variáveis após a poda: arrt vinho centrifugado e tempo de tratamento ácido. As variáveis aqui presentes não apresentam similaridade a com o modelo gerado anteriormente. O erro relativo de validação cruzada é de 68,38%, percebendo-se que a árvore gerada não representa bem os dados apresentados. A interpretação da árvore também não apresenta lógica frente ao que seria esperado, por exemplo, a presença de tempos de tratamento ácido maiores do que 2,5 (h) deveriam gerar deltas maiores, e não menores, o mesmo vale para arrt do vinho centrifugado em que valores maiores de açúcar residual da fermentação deveriam gerar deltas maiores, e não o oposto (Figura 15).

4. Conclusão

Apesar das abordagens diferentes entre si, podemos perceber que ambas explicaram os dados com base em variáveis comuns para as respostas de art do mosto e teor alcoólico. No entanto, em termos de aplicação, o uso de árvores de regressão provê informação mais intuitiva e fácil de ser interpretada. É interessante notar que para preservação das concentrações iniciais de açúcar no mosto o nível de impurezas minerais na cana deveria ter sido reduzido. Apesar da colheita mecanizada ter gerado um salto tecnológico para a área agrícola, sabe-se que esta provocou o aumento de impurezas minerais na matéria-prima. Na colheita manual, os níveis de impurezas de 2 a 5% eram típicos, enquanto que em colheita mecânica estes índices variam de 5 a 8%, podendo atingir índices de 10 a 20% em condições adversas [29]. Embora este trabalho não tenha tido objetivo preditivo, pode-se inferir que melhorias na área agrícola teriam beneficiado a qualidade do produto enviado para a etapa de fermentação. Outras variáveis de destaque são tempo de tratamento ácido e pH do fermento, ambas se referem a etapa de tratamento do levedo. Conforme verificado nas análises, tempos de tratamento ácido elevados foram prejudiciais às respostas de teor alcoólico, sendo válido conceber que a redução do tempo de tratamento ácido, < 2,3 (h), teria provido maiores respostas de etanol. Para as duas abordagens multivariadas utilizadas, este trabalho ainda indica que o tempo de espera de centrifugação não apresentou impacto na resposta de teor alcoólico, portanto o operador poderia ter optado por um tempo de espera maior a manter o levedo por mais tempo no tratamento ácido. Ainda sobre o tópico de tratamento ácido, embora garantir que o pH ideal seja atingido na etapa de tratamento ácido (indicado como 2,6 com o presente conjunto de dados) possa implicar em gastos maiores do consumível ácido sulfúrico, um balanço deve ser feito entre o etanol que deixa de ser produzido e o uso deste consumível. Por fim embora as variáveis temperatura do mosto e da fermentação não tenham aparecido nas árvores geradas, deveria ter sido considerada a necessidade de manter o processo controlado de forma a, respectivamente, evitar contaminação microbiana no mosto e manter a levedura em condições ótimas para conversão de açúcares em etanol. Embora note-se que a árvore gerada para variável delta não tenha convergido com o modelo, não sendo uma boa estratégia para representar esta variável, os outros modelos para açúcar total do mosto e teor alcoólico apresentaram uniformidades com as árvores geradas em suas respostas, indicando que os mesmos podem ser utilizados como complementares para explicar o comportamento dos dados apresentados.

Agradecimentos

Gostaria de agradecer aos professores do curso de Data Science e Big Data por tornarem esse trabalho possível, em especial a professor Walmes Zevianni pela orientação; à Novozymes Latin American LTDA pelo custeio e incentivo na execução do curso. Por fim, à Monique Oliveira, candidata ao doutorado em Engenharia Agrícola pela Unicamp, pelas discussões e suporte na execução deste trabalho.

Referências

[1] UNICA, 2018. http://www.unicadata.com.br/historico-de-producao-e-moagem.php?idMn=31&tipoHistorico=2

[2] MME, 2018. http://www.mme.gov.br/documents/1138769/0/P%26R+-+RenovaBio.pdf/a29044a3-6315-4845 -80d8-832852efbb7f

[3] M.O.S. Dias, R.M. Filho, P.E. Mantelatto, O. Cavalett, C.E.V. Rossell, A. Bonomi, M. R.L.V. Leal. *Sugarcane Processing for Ethanol and Sugar in Brazil*. Environmental Development 15, 35–51 (2015).

[4] S.R. Andrietta. *Modeling, Simulation and Control of Industrial Continuous Alcoholic Fermentation*, PhD thesis (FEA/Unicamp – Campinas-SP-Brazil, 1994).

[5] J. Fox and S. Weisberg.*Multivariate Linear Models in R*. An Appendix to An R Companion to Applied Regression*, 3rd ed (2018).

[6] L. Breiman, J. Friedman and C.J. Stone. *Classification and Regression*, Taylor & Francis, (1984).

[7] G. De'ath and K.E Fabricius. *Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis*. Ecology, 81, 3178– 3192 (2000).

[8] N. Speybroeck. *Classification and regression trees*. Int J Public Health 57, 243–246 (2012).

[9] R Development Core Team. (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, url http://www.R-project.org.

[10] M. Gagolewski, B. Tartanus and other contributors (stringi source code), IBM, Unicode, Inc., other contributors (ICU4C source code); Unicode, Inc. (Unicode Character Database). *Character String Processing Facilities*. https://cran.r-project.org/web/packages/stringi/stringi.pdf. (2019).

[11] H. Wickham and RStudio. *Simple, Consistent Wrappers for Common String Operations*. https://cran.r-project.org/web/pack ages/stringr/stringr.pdf. (2019).

[12] H. Wickham, L. Henry, RStudio. *Easily Tidy Data with 'spread()' and 'gather()' Functions*. https://cran.rproject.org/web/packages/tidyr/tidyr.pdf.

[13] W. Borzani. *Batch Ethanol Fermentation: The Correlation between the Fermentation Efficiency and the Biomass Initial Concentration Depends on What is Considered as Produced Ethanol*. Brazilian Journal of Microbiology 37, 87-89 (2006).

[14] T. Wei, V. Simko, M. Levy, Y. Xie, Y. Jin, J. Zemla. Visualization of a Correlation Matrix. https://cran.r-

project.org/web/packages/corrplot/corrplot.pdf. (2017)

[15] A. Hebbali. *Tools for Building OLS Regression Models*. https://cran.r-project.org/web/packages/olsrr/olsrr.p df, (2018).

[16] J. Fox, S. Weisberg, B. Price, D. Adler, D. Bates, G. Baud-Bovy, B. Bolker, S. Ellison, D. Firth, M. Friendly, G. Gorjanc, S. Graves, R. Heiberger, R. Laboissiere, M. Maechler, G. Monette, D. Murdoch, H. Nilsson, D. Ogle, B. Ripley, W. Venables, S. Walker, D. Winsemius, A. Zeileis, R-Core. *Companion to Applied Regression*. https://cr an.r-project.org/web/packages/car/car.pdf, (2019).

[17] O. Valsechi. *Microbiologia em Açúcares de Cana-de-Açúcar*. Piracicaba:ESALQ/USP (2000).

[18] T. F. Z. Brassolatti, R. C. Vieira, M. A. B. Costa, M. Brassolatti. *Análise do Percentual de Impurezas Vegetais e Minerais Presentes na Cana-de-Açúcar*. Revista Interdisciplinar de Tecnologias e Educação (2016).

[19] U. Farooq, F.M. Anjum, T. Zahoor, Sajjad-Ur-Rahman, M.A. Randhawa, A. Ahmed, K. Akram. *Optimization of Lactic Acid Production from Cheap Raw Material: Sugarcane Molasses*. Pak. J. Bot., 44:333-338, (2012).

[20] D.I.P. Atala. Fermentação Alcoólica com Alta Densidade Celular: Modelagem Cinética, Convalidação de Parametros e Otimização Do Processo. Master thesis (FEA/Unicamp – Campinas- SP-Brazil, 2000).

[21] H.V. de Amorim, L.C. Basso, D.M.G Alves. *Processos de Produção de Álcool* (Centro de Biotecnologia Agrícola, Piracicaba, 1996).

[22] Melo H.F. *Resposta ao Estresse Ácido em Leveduras de Fermentação Alcoólica Industrial.* Tese de Doutorado (Departamento de Micologia/ UFPE, 2006).

[23] C. Dorta, P. Oliva-Neto, M.S. de-Abreu-Neto, N. Nicolau-Junior, A.I. Nagashima. *Synergism among lactic acid, sulfite, pH and ethanol in alcoholic fermentation of Saccharomyces cerevisiae (PE-2 and M-26)*. World Journal of Microbiology & Biotechnology, 22: 177–182 (2006).

[24] V. Carmelo, R. Santos, C.A. Viegas, I. Sá-Correio. *Modification of Saccharomyces cerevisiae thermotolerance following rapid exposure to acid stress*. International Journal of Food Microbiology (1998).

[25] F.L.C. Mensonides, J.M. Schurmans, M.J.T. Mattos, K.J. Hellingwers, S. Brul. *The metabolic response of Saccharomyces cerevisiae to continuous heat stress*. Molecular Biology Reports (2002).

[26] H.V. Amorim, M.L. Lopes, J.V.C. Oliveira, M.S. Buckeridge, G.H.Goldman. *Scientific challenges of bioethanol production in Brazil*. Appl Microbiol Biotechnol 91:1267–1275 (2011).

[27] T. Therneau, B. Atkinson, B. Ripley. *Recursive Partitioning and Regression Trees*. https://cran.r-project.org/web/packages/rpart/ rpart.pdf. (2019).

[28] S. Milborrow. *Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'*. https://cran.r-project.org/web /pack ages/rpart.plot/rpart.plot.pdf. (2019).

[29] P.G. Magalhães. *Qualidade da Matéria-Prima Entregue nas Usinas*. Workshop sobre Produção de Etanol: Qualidade da Matéria-Prima. Lorena, (2008).

Apêndice

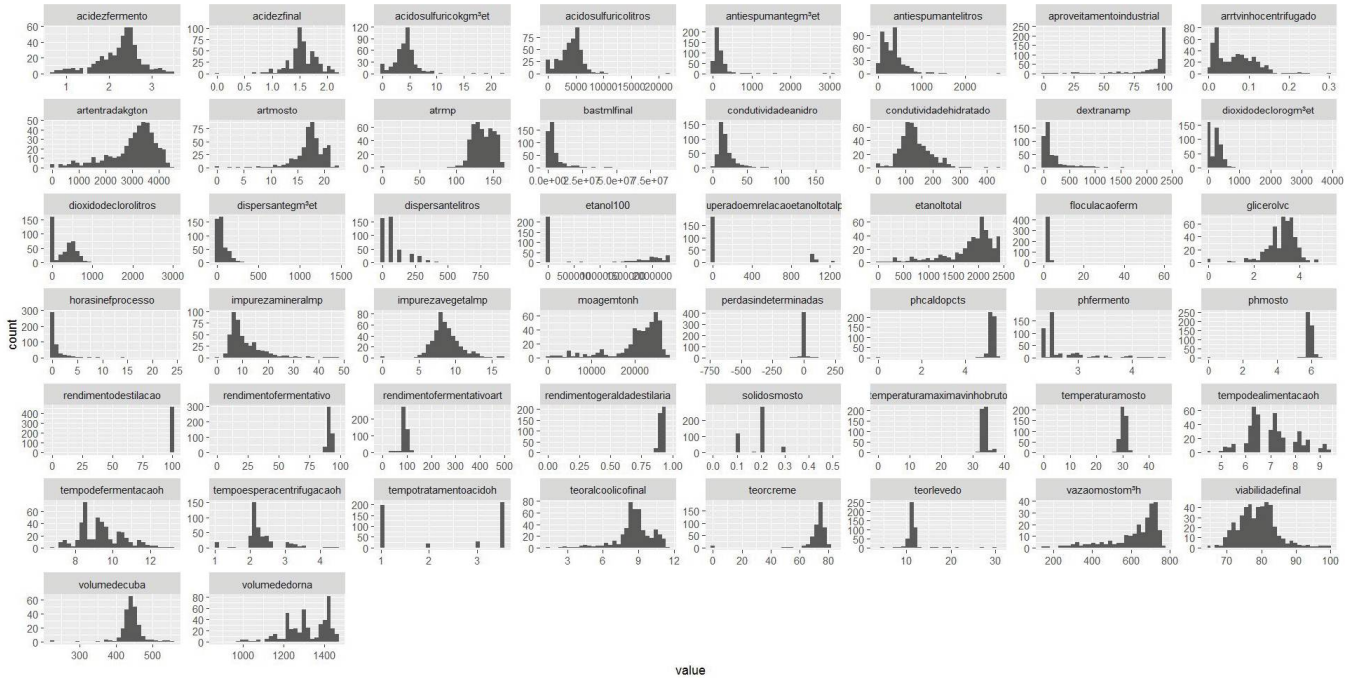

Figura 1. Histograma de frequência das diferentes variáveis do processo de produção de etanol a partir de cana-de-açúcar.

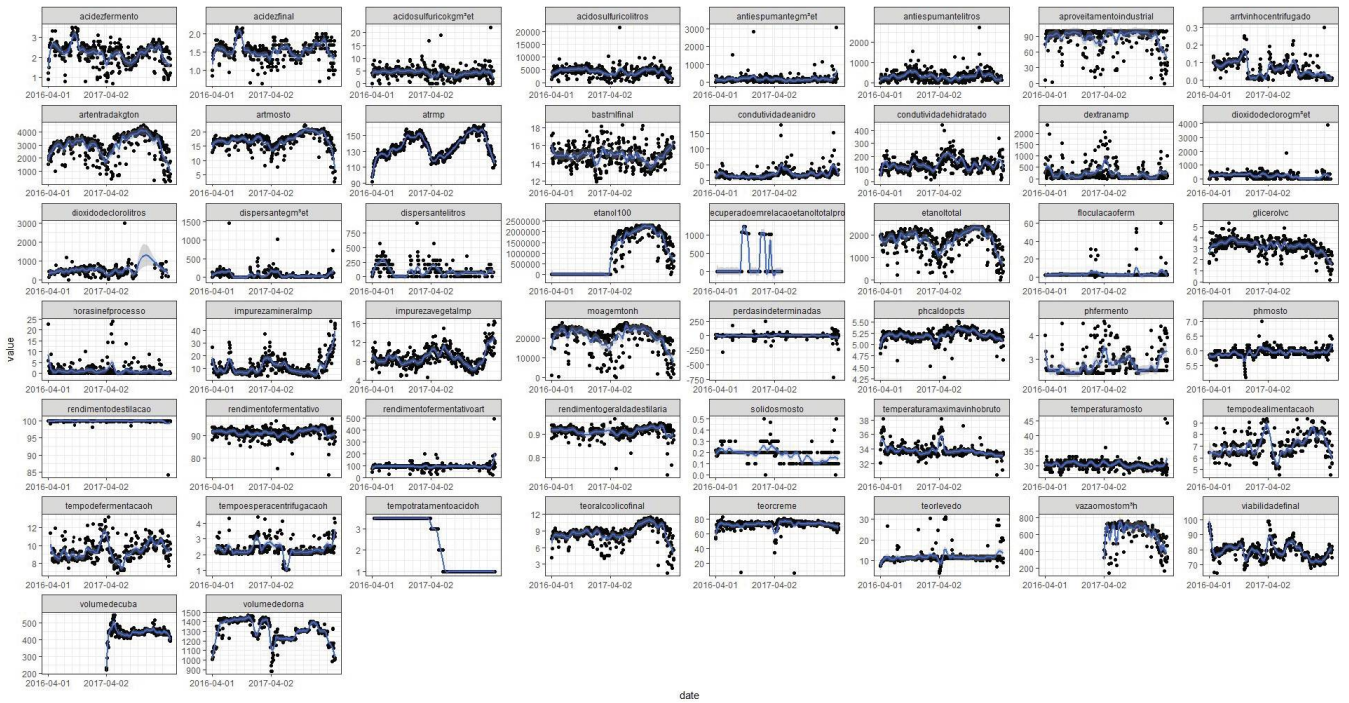

Figura 2. Séries temporais das diferentes variáveis do processo de produção de etanol a partir de cana-de-açúcar após aplicação de filtros.

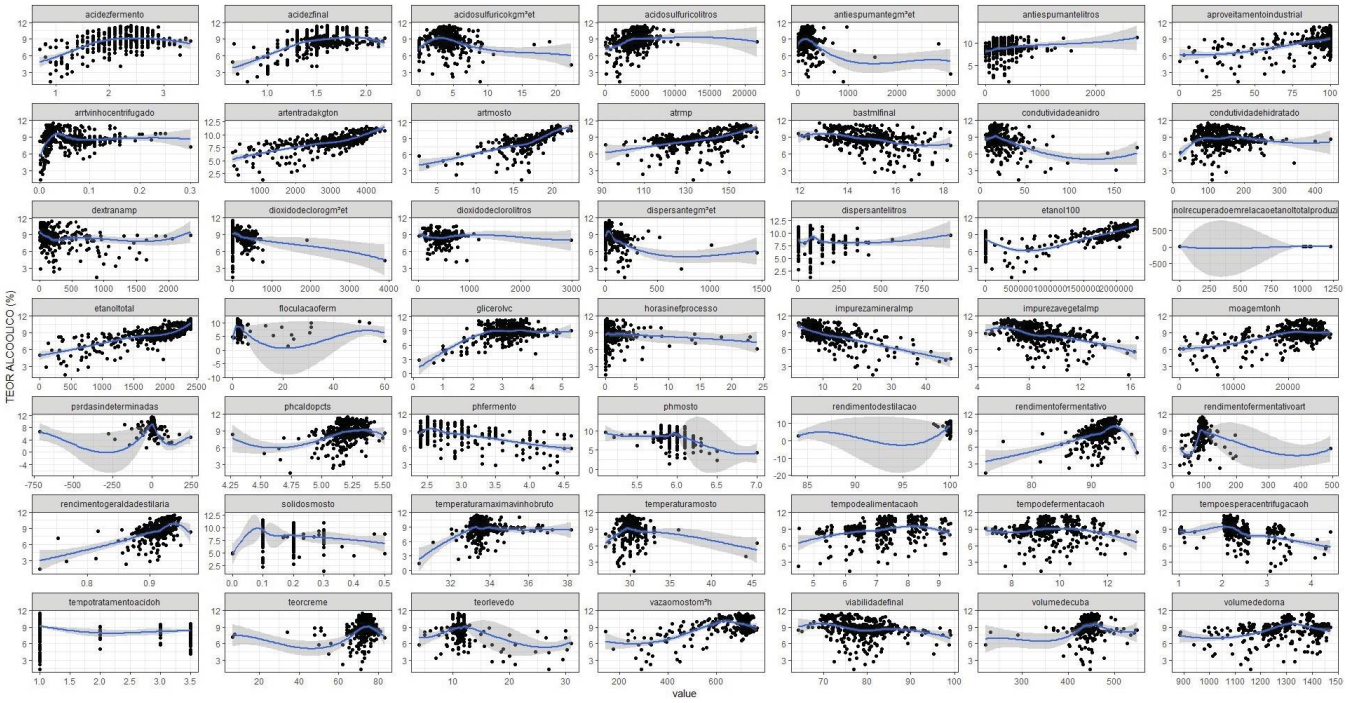

Figura 3. Gráficos do teor alcoólico do vinho bruto frente as diferentes variáveis do processo de produção de etanol.

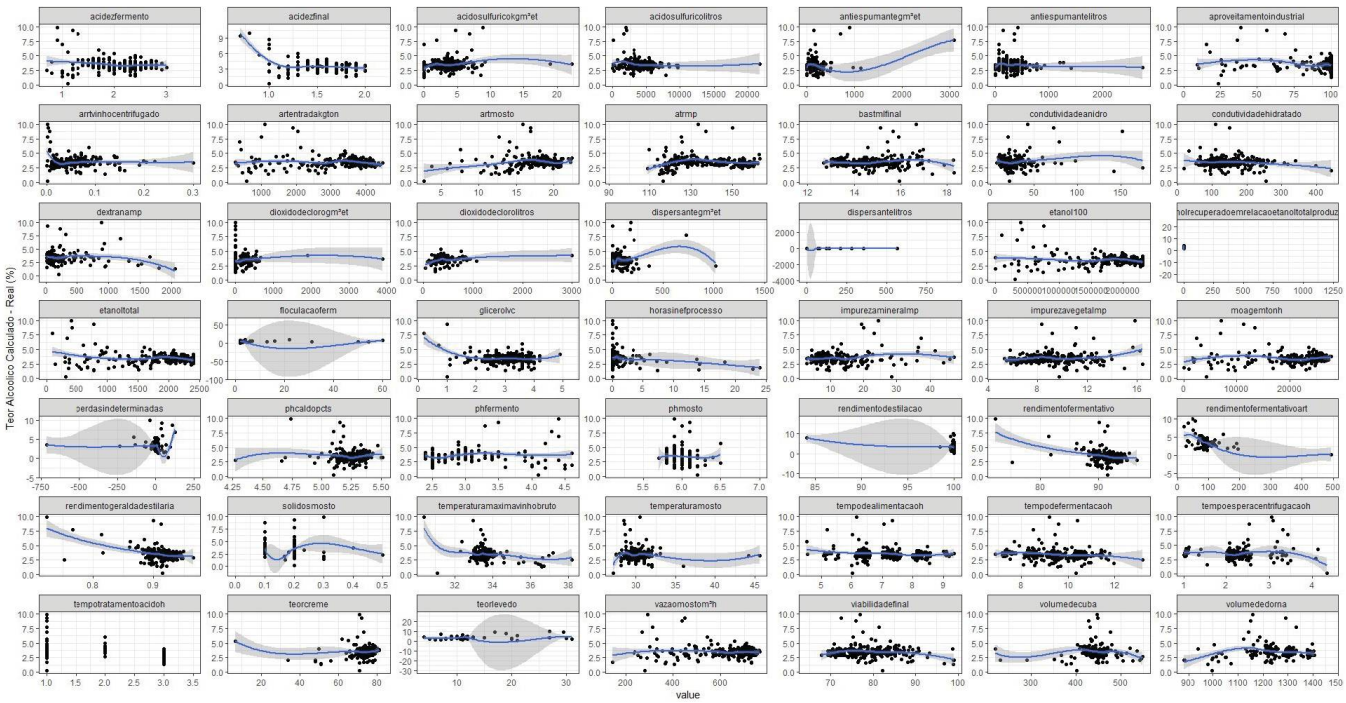

Figura 4. Gráficos do teor alcoólico real – calculado (delta) frente as diferentes variáveis do processo de produção de etanol.

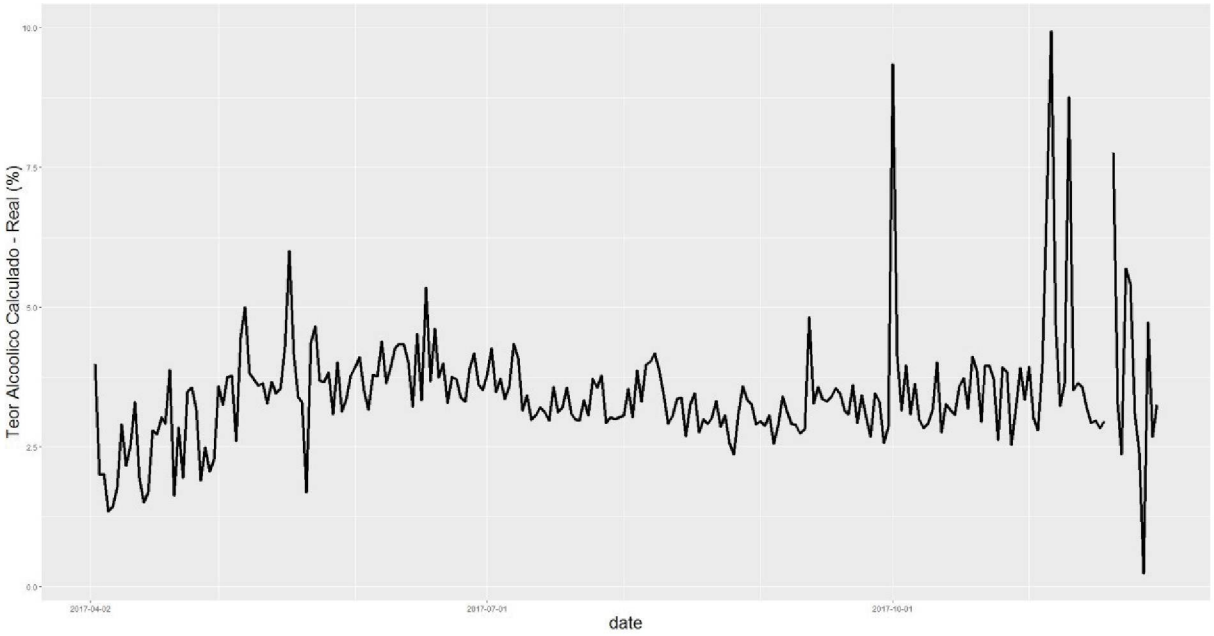

Figura 5. Teor alcoólico calculado – real ao longo das duas safras.

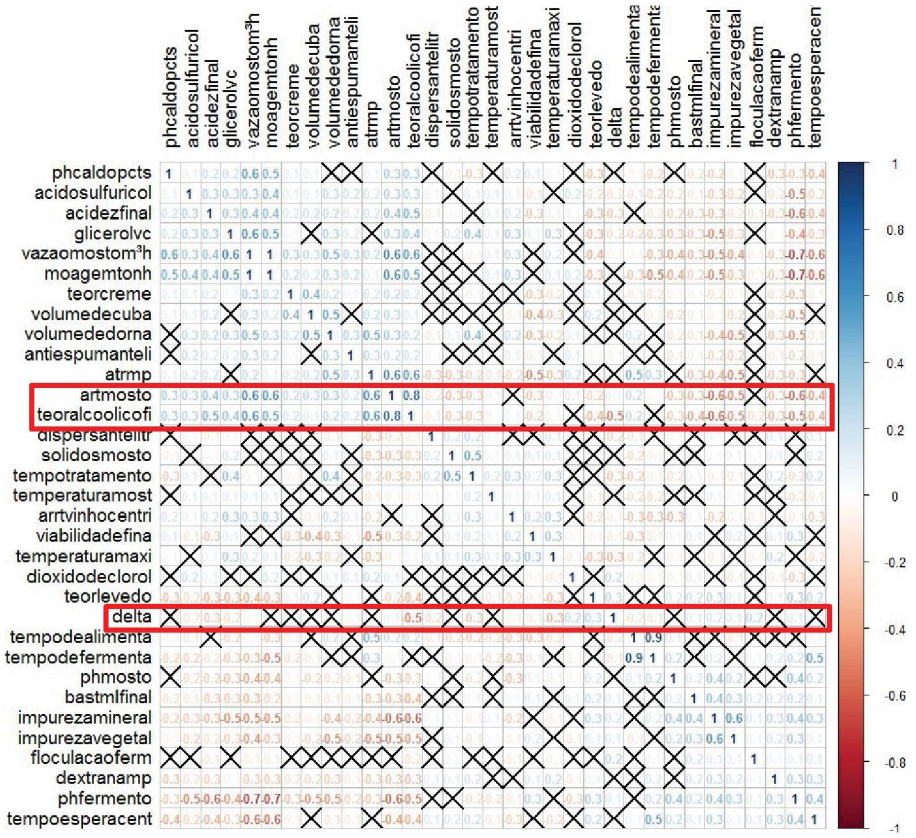

Figura 6. Matriz de correlação entre as diferentes variáveis do processo de produção de etanol a partir de cana-de-açúcar em que a escala gradual de cores denota a intensidade da correlação sendo azul positiva e vermelho negativa.

Model Summary

R	0.848	RMSE	0.071
R-Squared	0.719	Coef. Var	9.576
Adj. R-Squared	0.716	MSE	0.005
Pred R-Squared	0.705	MAE	0.047

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	5.861	5	1.172	231.169	0.0000
Residual	2.287	451	0.005		
Total	8.148	456			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	0.394	0.027		14.333	0.000	0.340	0.448
atrmp	0.330	0.020	0.477	16.922	0.000	0.292	0.369
moagemtonh	0.314	0.019	0.482	16.574	0.000	0.277	0.351
impurezamineralmp	-0.135	0.026	-0.158	-5.159	0.000	-0.186	-0.084
solidosmosto	-0.144	0.028	-0.135	-5.160	0.000	-0.199	-0.089
temperaturamosto	-0.090	0.043	-0.055	-2.124	0.034	-0.174	-0.007

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	atrmp	addition	0.385	0.384	534.7870	-759.7556	0.1049
2	moagemtonh	addition	0.683	0.681	58.7750	-1060.0602	0.0755
3	impurezamineralmp	addition	0.700	0.698	33.6180	-1083.0786	0.0735
4	solidosmosto	addition	0.717	0.714	8.5130	-1107.4935	0.0715
5	temperaturamosto	addition	0.719	0.716	6.0000	-1110.0436	0.0712

Figura 7. Sumário do modelo, teste de ANOVA, estimativa dos parâmetros e sumarização da seleção de variáveis por meio do método AIC para o modelo do açúcar redutor total (art) do mosto referente aos dados da safra de 2016-2017.

Model Summary

R	0.850	RMSE	0.074
R-Squared	0.722	Coef. Var	10.171
Adj. R-Squared	0.717	MSE	0.005
Pred R-Squared	0.692	MAE	0.053

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	6.198	9	0.689	125.535	0.0000
Residual	2.381	434	0.005		
Total	8.578	443			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	0.456	0.046		9.873	0.000	0.365	0.547
impurezamineralmp	-0.259	0.029	-0.286	-8.991	0.000	-0.316	-0.203
atrmp	0.297	0.023	0.395	12.984	0.000	0.252	0.341
acidezfinal	0.065	0.028	0.072	2.290	0.022	0.009	0.120
tempotratamentoacidoh	-0.116	0.009	-0.402	-13.614	0.000	-0.133	-0.100
temperaturamaximavinhobruto	0.242	0.040	0.173	5.991	0.000	0.162	0.321
phfermento	-0.134	0.022	-0.199	-6.011	0.000	-0.178	-0.090
glicerolvc	0.174	0.039	0.155	4.515	0.000	0.098	0.250
teorlevedo	-0.114	0.051	-0.063	-2.242	0.025	-0.214	-0.014
floculacaoferm	-0.097	0.048	-0.052	-2.031	0.043	-0.191	-0.003

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	impurezamineralmp	addition	0.367	0.365	810.0390	-642.0439	0.1199
2	atrmp	addition	0.510	0.508	525.5060	-758.0801	0.1056
3	acidezfinal	addition	0.577	0.574	394.3380	-821.0318	0.0983
4	tempotratamentoacidoh	addition	0.637	0.634	264.0780	-876.1070	0.0907
5	temperaturamaximavinhobruto	addition	0.673	0.669	156.5010	-940.6903	0.0835
6	phfermento	addition	0.711	0.707	90.4540	-993.2962	0.0787
7	glicerolvc	addition	0.717	0.712	39.0640	-1034.5338	0.0746
8	teorlevedo	addition	0.720	0.715	36.1780	-1037.2030	0.0743
9	floculacaoferm	addition	0.722	0.717	33.8260	-1039.4038	0.0741

Figura 8. Sumário do modelo, teste de ANOVA, estimativa dos parâmetros e sumarização da seleção de variáveis por meio do método AIC para o modelo do teor alcoólico referente aos dados da safra de 2016-2017.

Model Summary

R	0.396	RMSE	0.094
R-Squared	0.157	Coef. Var	28.664
Adj. R-Squared	0.150	MSE	0.009
Pred R-Squared	0.082	MAE	0.062

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	0.390	2	0.195	21.921	0.0000
Residual	2.099	236	0.009		
Total	2.489	238			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	0.548	0.034		16.119	0.000	0.481	0.615
temperaturamaximavinhobruto	-0.310	0.058	-0.320	-5.347	0.000	-0.424	-0.196
acidezfinal	-0.152	0.039	-0.231	-3.856	0.000	-0.229	-0.074

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	temperaturamaximavinhobruto	addition	0.103	0.100	211.5840	-435.2438	0.0969
2	acidezfinal	addition	0.157	0.150	187.4200	-445.3714	0.0943

Figura 9. Sumário do modelo, teste de ANOVA, estimativa dos parâmetros e sumarização da seleção de variáveis por meio do método AIC para o modelo do delta do teor alcoólico (calculado – real) para referente aos dados da safra de 2017.

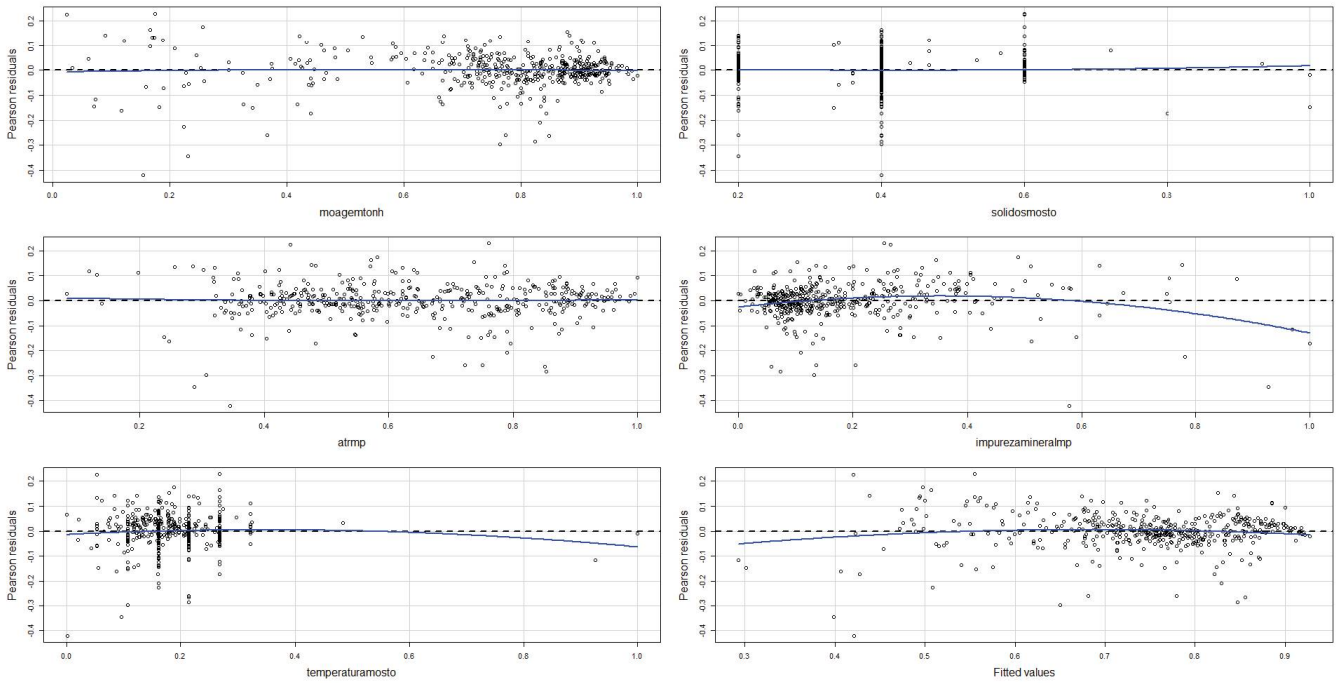

Figura 10. Resíduos marginais da resposta de ART do mosto frente as diferentes variáveis explicativas e do modelo normaliza dos.

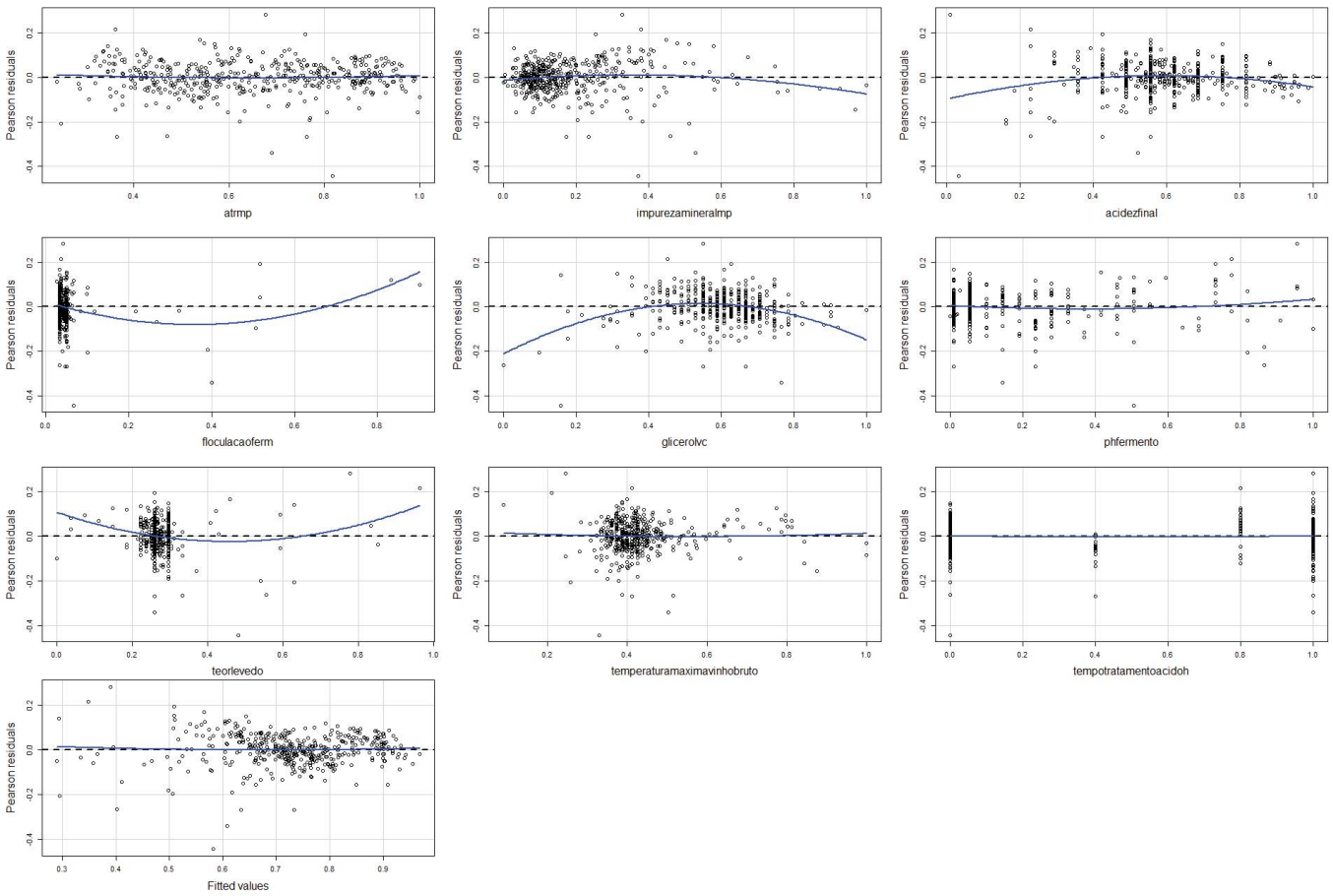

Figura 11. Resíduos marginais da resposta de teor alcoólico frente as diferentes variáveis explicativas e do modelo normalizados.

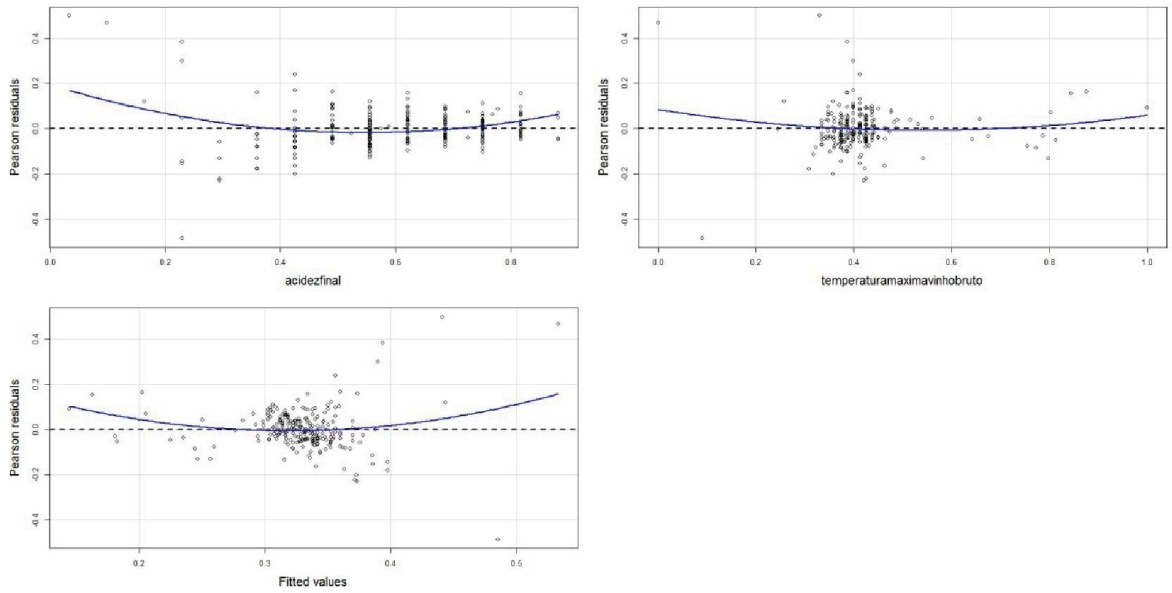

Figura 12. Resíduos marginais da resposta de delta (teor álcoolico calculado – real) frente as diferentes variáveis explicativas e do modelo normalizados.

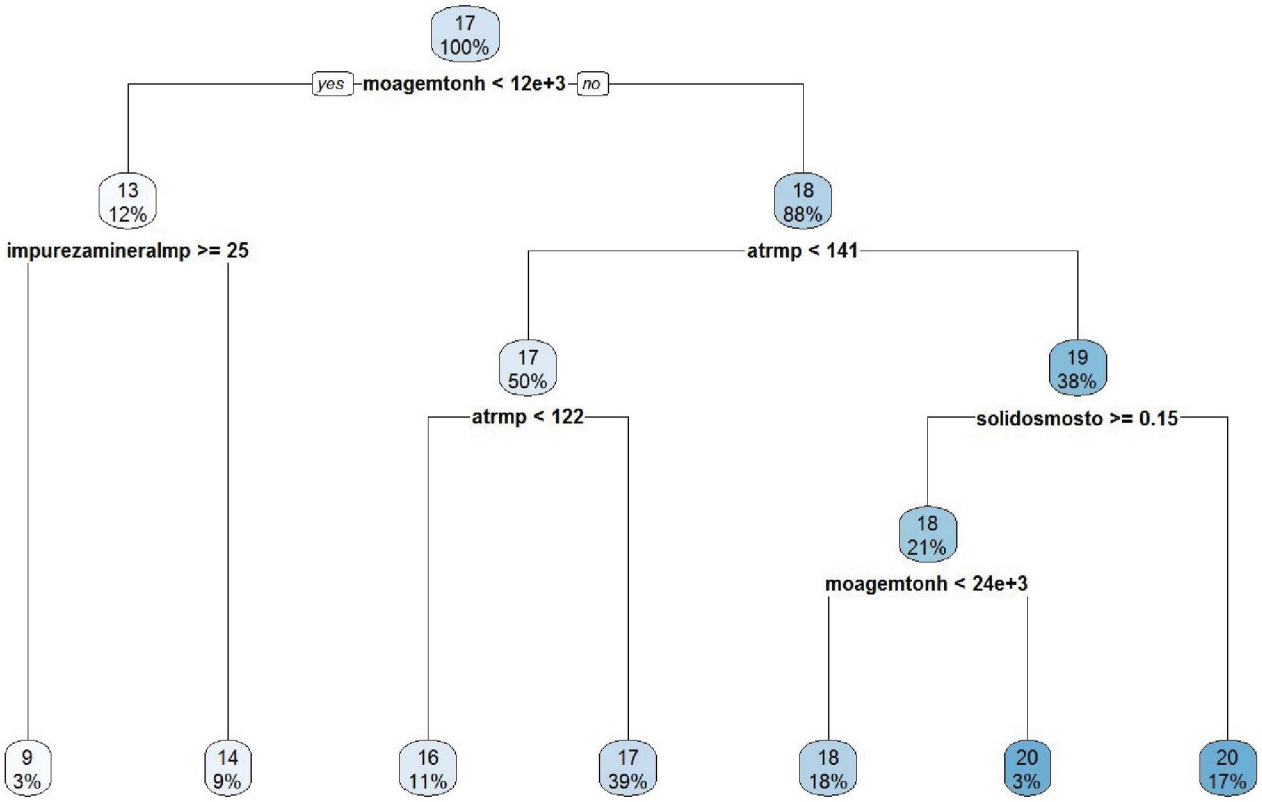

Figura 13. Árvore de decisão para a resposta de açúcar redutor do mosto (art do mosto).

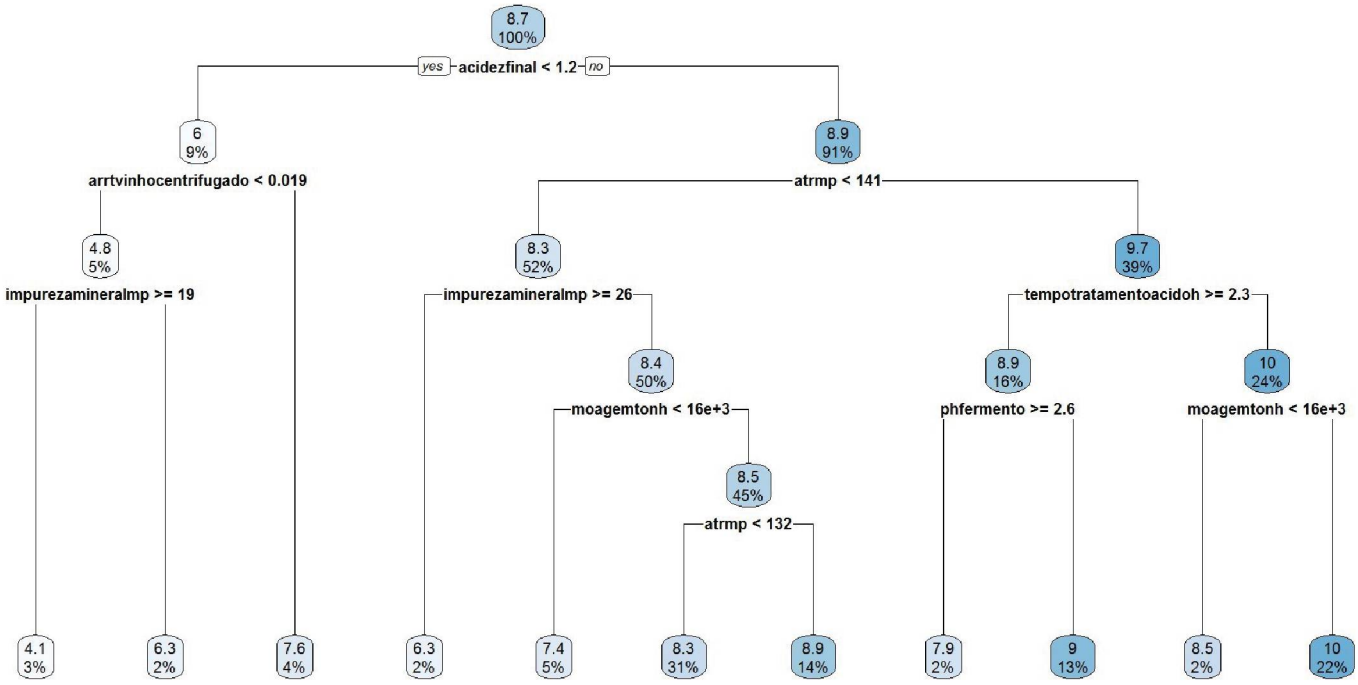

Figura 14. Árvore de decisão para a resposta de teor alcoólico do vinho bruto.

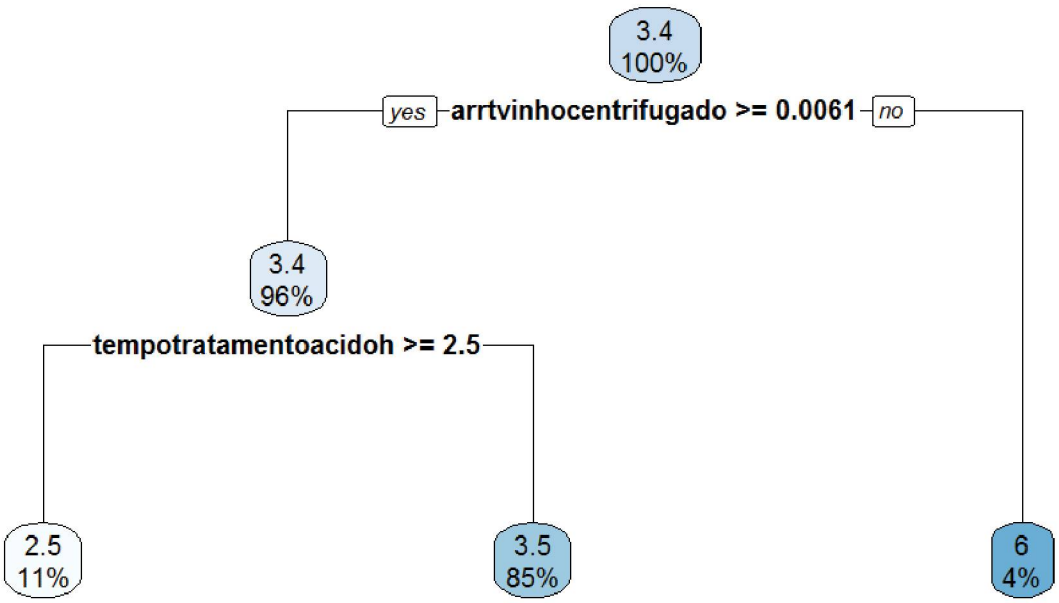

Figura 15. Árvore de decisão para a resposta de teor alcoólico calculado – real (delta).