Análise Causal com Dados da Embrapa

1. Introdução

Este projeto tem como objetivo analisar dados de plantio da Embrapa, com variáveis como localização,espécie,fase de crescimento,aditivo e área. Os frutos foram analisados sob diversos aspectos,como produtividade,quantidade de frutose,sacarose recuperável,etc. Obviamente, se o interesse fosse determinar as melhores características de plantio para se obter algum resultado específico no fruto seria preciso desenvolver um experimento,com o cuidadoso balanceamento das variáveis, sendo que para isto existe ampla literatura em estatística.Analisando dados observacionais como foi feito aqui as conclusões são mais restritas daquelas que obteríamos com um experimento, mas o objetivo aqui foi obter o máximo de conclusões a partir de um 'experimento incompleto'.

2. Características dos dados

O conjunto de dados possui 7653 linhas e 16 colunas,listadas abaixo com sua descrição:

REGIAO - local de plantio(11 regiões)

ESPÉCIE - espécie de planta(148 especies)

FASE - fase de crescimento(9 fases)

ADITIVO - aditivo usado(6 aditivos)

AREA - área de plantio

ABA_ESP - Quantidade de Frutose

ABB_ESP - Teor de sacarose aparente

ABC_ESP - Quantidade de Fibra da fruta

ABD_ESP - Tonelada de frutose por hectare

ABE_ESP - Produtividade

ABF_ESP - Quantidade de Caldo

ABG_ESP - Pureza da Frutose

ABH_ESP - Idade da espécie

ABI_ESP - Sacarose total recuperável

ABL_ESP - Toneladas de ABI_ESP por hectare

ABM_ESP - Quantidade de Frutose redutora

Seria realmente desafiador criar um experimento com um número tão grande de espécies.

3. Abordagem Utilizada

Para resolver este problema foi utilizada a biblioteca cfml_tools, de machine learning contrafactual: https://github.com/gdmarmerola/cfml_tools

O clustering supervisionado desta biblioteca verifica como as mudanças na variável de tratamento refletem nas mudanças no alvo, dados os clusters determinados pelas variáveis explicativas que mais impactam o alvo.

4. Resultados do modelo

Os resultados para a variável Quantidade de Frutose são mostrados abaixo. É possível notar que a análise só utiliza apenas 3 regiões, aproximadamente 10 espécies e 2 fases.

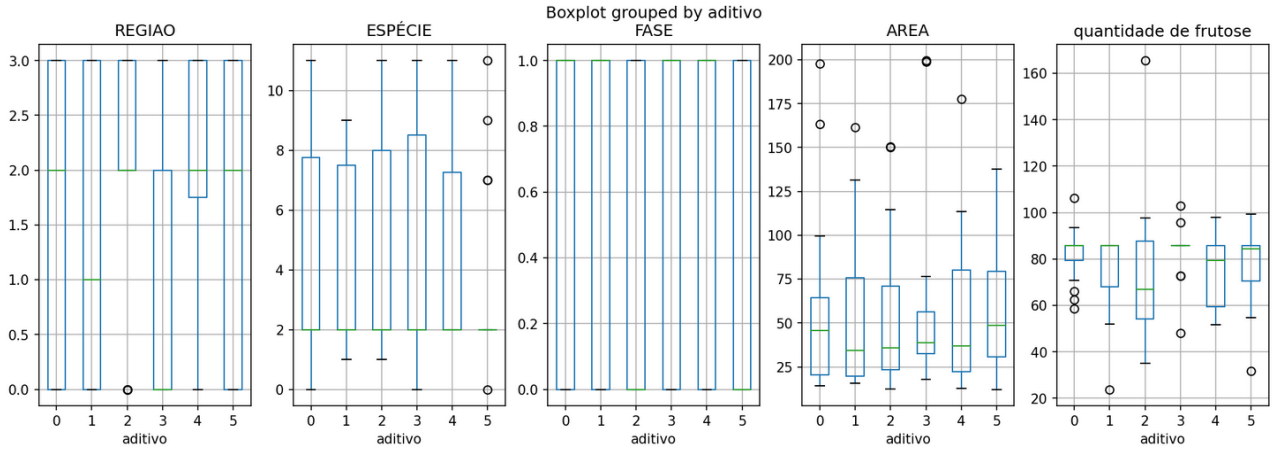

Fig1-Resultados para quantidade de frutose

Aditivo	0	1	2	3	4	5
Media de frutose	82.4	73.7	72.2	83.0	73.9	76.8
Desvio padrão	10.7	20.6	26.7	12.5	14.8	15.1

Tab1- média e desvio padrão da frutose

É possível observar que para as regiões de São Carlos(0), Assis(1), Minas Gerais(2) e Araçatuba(3),para as fases 5d(0) e >5d(1) o aditivo B(0) gerou o melhor resultado de frutose em 10 espécies, com um bom resultado de média,com valor 82.4, e menor desvio padrão de 10.7. Para outras regiões não é possível afirmar que tipo de aditivo gera melhores resultados, pois os dados existentes não nos permitem estas conclusões.As 10 espécies para as quais estas conclusões são válidas: SCS16', 'PS321832847', 'SCS14', 'PS321716180', 'BULL24825336','BULL24785554', 'PS321841431', 'PS321801842', 'SCS115','BULL24855536', 'BULL24845210', 'PS321813250'.

Os resultados para a variável Sacarose Total Recuperável são mostrados abaixo. É possível notar que a análise só utiliza apenas uma região, aproximadamente 30 espécies e 2 fases.

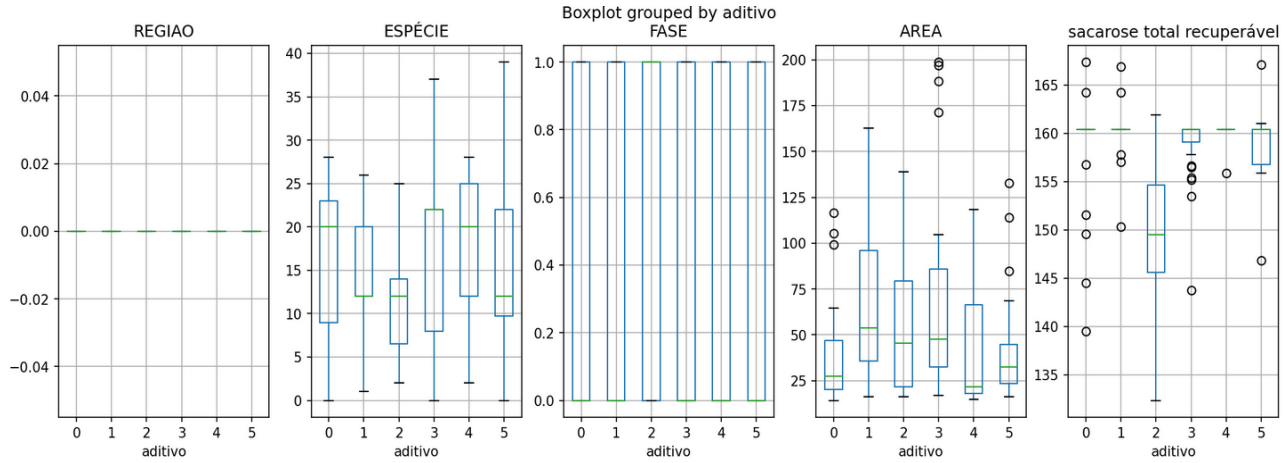

Fig2-Resultados para sacarose total recuperável

Aditivo	0	1	2	3	4	5
Media de sacarose	158.5	160.0	149.0	158.7	159.8	159.0
Desvio padrão	5.7	3.2	9.5	3.6	1.4	3.6

Tab2- média e desvio padrão da sacarose

Aqui vemos que para a região de São Carlos(0),para as fases 5d(0) e >5d(1) o aditivo D(4) gerou o melhor resultado de sacarose recuperável em 28 espécies, com um bom resultado de média,com valor 159.8, e menor desvio padrão de 1.4. Para outras regiões não é possível afirmar que tipo de aditivo gera melhores resultados, pois os dados existentes não nos permitem estas conclusões.As 28 espécies para as quais estas conclusões são válidas: 'SCS16', 'PS321832847', 'SCS14', 'PS321716180', 'BULL24825336','BULL24785554', 'PS321841431', 'PS321801842', 'SCS115','BULL24855536', 'BULL24845210', 'PS321813250', 'BULL24867515','SCS12', 'SCS117', 'SCS19', 'SCS122', 'PT35963346', 'SCS114','PS321801816', 'BULL24855156', 'BULL24855453', 'PS321803280','SCS19005HP', 'DIVERSAS', 'BULL2492579', 'SCS120', 'AIC456911099','AIC456PS321955000'.

Aqui são mostrados os resultados para a soma normalizada de todas as variáveis resposta: Quantidade de Frutose,Teor de sacarose aparente,Quantidade de Fibra da fruta,Tonelada de frutose por hectare,Produtividade,Quantidade de Caldo,Pureza da Frutose,Idade da espécie, Sacarose total recuperável,Toneladas de ABI_ESP por hectare,Quantidade de Frutose redutora.

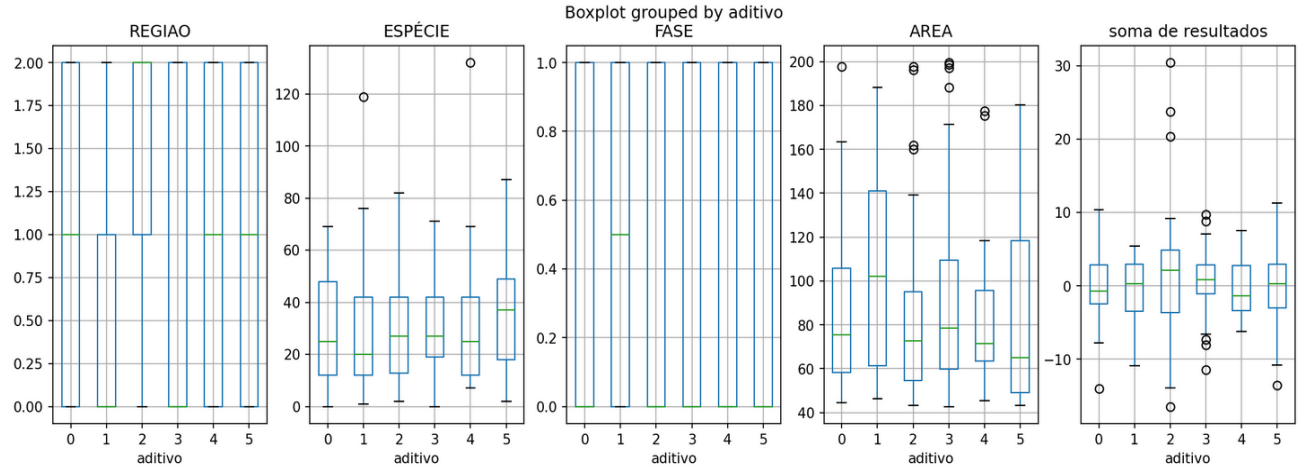

Fig3-Resultados para soma normalizada

Aditivo	0	1	2	3	4	5
Media de soma norm.	-0.4	-0.3	1.4	0.5	-0.5	-0.02
Desvio padrão	4.4	4.1	9.0	4.3	3.8	5.5

Tab3- média e desvio padrão da soma normalizada

Para as regiões de São Carlos(0), Assis(1), Minas Gerais(2),com as fases 5d(0) e >5d(1), nenhum aditivo gerou diferenças significativas de soma dos resultados em aproximadamente 50 espécies,como pode ser observado no último grafico boxplot,soma de resultados. O Aditivo 2 gerou maior média de 1.4, mas maior desvio padrão de 9. O aditivo 3 alcançou o melhor resultado combinado de média e desvio padrão(media 0.5 e desvio 4.3),mas mesmo assim não tem um diferença tão grande em relação aos outros aditivos.

5. Conclusão

Este trabalho mostrou que é possível uma inferência causal através de dados observacionais. Os dados são desafiadores por contarem com um número muito grande de espécies(148), utilizadas em 11 regiões. Apesar das inferências realizadas em apenas 3 das 11 regiões, algumas conclusões puderam ser feitas.

6. Referências

1. CFML-tools: https://github.com/gdmarmerola/cfml_tools