RELATÓRIO DA BASE DE DADOS UNIFICADA E GUIA DE BOAS PRÁTICAS PARA COLETA DE DADOS

Ana Letícia Becker Gomes – Universidade do Vale do Itajaí
Anita Maria da Rocha Fernandes – Universidade do Vale do Itajaí
Daniel Pereira Guimarães – Empresa Brasileira de Pesquisa Agropecuária
Fábio Volkmann Coelho – Universidade do Vale do Itajaí
Maurílio Fernandes de Oliveira – Empresa Brasileira de Pesquisa Agropecuária
Ramon Costa Alvarenga – Empresa Brasileira de Pesquisa Agropecuária

Itajaí, julho de 2025

O presente documento tem por objetivo explicar em detalhes a base unificada de dados, elaborada durante a realização do projeto que se deu como parceria entre a Universidade do Vale do Itajaí (UNIVALI) e a Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) da Unidade Milho e Sorgo, o qual foi contemplada pela Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC) nos projetos da chamada pública FAPESC N° 54/2022 e também pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) no projeto da chamada CNPq/MCTI No 10/2023, faixa A, edital universal, processo N° 404755/2023-2.

Tal relatório será entregue a EMBRAPA como produto técnico resultante da dissertação intitulada "*Modelos de Aprendizado de Máquina para Predição de Dinâmicas Populacionais de Plantas Daninhas em Sistemas ILP*", da estudante de mestrado Ana Letícia Becker Gomes, que conduziu os experimentos iniciais do projeto citado acima.

Além da descrição da base de dados unificada, tem-se também que tal documento apresenta um guia de boas práticas para coleta de dados, o qual visa auxiliar os profissionais da EMBRAPA durante as amostragens e – por consequência – contribuir para o atingimento de melhores resultados do projeto.

1. BASES DE DADOS

Tem-se que o objetivo geral do projeto – e da dissertação feita com base neste – era o desenvolvimento de modelos de aprendizado de máquina para predição de culturas e épocas em que haveria a emergência de plantas daninhas em Sistemas Integração Lavoura-Pecuária (ILP).

Assim, tem-se que uma base de dados unificada foi desenvolvida a partir da junção de outras três bases de dados, as quais contém informações sobre os três elementos investigados durante o desenvolvimento do trabalho: as espécies de plantas daninhas, o solo da região, e também o clima do local.

As bases individuais também foram produzidas pela estudante de mestrado, a partir dos dados fornecidos pela EMBRAPA. Logo, as seções 1.1, 1.2 e 1.3 apresentam as especificações de cada uma destas bases.

1.1. Base de dados das plantas daninhas

- **Fonte dos dados**: EMBRAPA Milho e Sorgo
- **Formato**: planilhas EXCEL (.xlsx)
- **Total de registros:** 1104
- **Anos de coleta**: 2006, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023 e 2024
- **Variáveis**:
    - **Data**: refere-se ao dia em que determinada espécie de planta daninha foi coletada. Os possíveis valores que tal variável pode assumir estão dentro do período mencionado acima.
    - **Invasora:** nome comum da espécie que foi coletada.
        - Valores: *A. retroflexus*; Angiquinho; Apaga-Fogo; Assa-Peixe; Beldroega; Botão de Ouro; Buva; Cabreuva; Capim; Capim Amargoso; Capim Braquiária; Capim Carrapicho; Capim Colchão; Capim Custódio; Capim Favorito; Capim Guiné; Capim Marmelada; Capim

Mombaça; Carrapicho de Carneiro; Carrapicho Rasteiro; Carrapicho Redondo; Caruru; Chic-Chic; Cipó; Corda de Viola; Cordão de Frade; Crista de Galo; Crucífera; Erva de Santa Luzia; Erva de Touro; Erva Moura; Erva Quente; Fedegoso; Grama Seda; Guaxuma; Joá de Capote; Leiteiro; Lobeira; Losna Branca; Macela; Malícia; Malva Branca; Malva Preta; *Malvastrum*; Mamona; Maxixe; Mentrasto; Pé-de-Galinha; Picão; Poaia; Serralha; Sida; *Sidastrum;* Soja Perene; Sorgo Selvagem; Tiguera de Milho; Timbete; Tiririca; Trapoeraba; Vassoura de Bruxa; Vassoura Rabo de Tatu; *Xanthium stra.*

- **Tipo de Folha:** diz respeito à morfologia da folha da planta daninha coletada.
  - Valores: Folha Larga (dicotiledônea) ou Folha Estreita (monocotiledônea).
- **Quantidade**: número de espécies coletadas de uma determinada espécie de planta daninha. Tal variável é numérica.
- **Peso Verde (g):** refere-se a soma dos pesos de todas as plantas daninhas coletadas de uma determinada espécie no dia da amostragem (biomassa verde). É uma variável numérica.

Observação: para ter um maior volume de dados, também foram adicionados registros sobre as plantas daninhas de caminhamento. Isto é, durante o percurso até os pontos de coleta, foram anotadas as espécies que apareceram no caminho. Nestes casos, foi considerada uma planta de cada espécie vista. Também foi definido que o peso verde seria 1g e o peso seco 0,1g (mais detalhes sobre o peso seco no próximo item).

Observação: em outros casos, apesar de não haver informação sobre o peso verde, havia os valores do peso seco. Nestes casos, foi utilizado uma regra de três para descobrir o valor do peso verde a partir do peso seco, já que o segundo corresponde a – aproximadamente – 40% do primeiro, conforme foi explicado pelo agrônomo do projeto.

- **Peso Seco (g):** após a pesagem para averiguar o Peso Verde, as plantas daninhas passam por um processo de desidratação. Depois deste procedimento, o peso total de todas plantas de uma mesma espécie coletada é

novamente verificado, o qual corresponde ao Peso Seco. Esta também é uma variável numérica.

Observação: para os registros de 2006 e também para alguns de 2015, os valores do Peso Seco estavam faltando. Para poder manter estes dados, foi optado por preencher tais valores faltantes pelo valor da mediana. Para isso – com base nos valores dos outros registros – o peso seco individual de cada espécie foi calculado. A partir destes, foi calculada a mediana do peso seco de cada uma destas espécies. Por último, tal valor foi aplicado nos registros faltantes.

Observação: durante este processo, percebeu-se que algumas espécies possuíam apenas um registro na base, de modo que – por não haver mais registros – não foi possível calcular a mediana do peso seco para preencher as lacunas faltantes. Assim, tem-se que estes registros foram excluídos; são estas as espécies: Vidoeiro Branco; Carrapicho Beiço de Boi e Erva Palha.

- **Coleta da Amostra:** remete à época de amostragem das plantas daninhas.
  - Valores:
    - Entre-Safra: é o período entre uma lavoura e outra. Há a presença de gado durante este tempo.
    - Na Colheita: é o momento antes ou imediatamente após a colheita.
    - Na Lavoura: é a época logo depois do plantio.

Observação: inicialmente esta variável também podia assumir o valor "Pré-Dessecação", que é um período específico da entre-safra, todavia, foi optado – por orientação do agrônomo da EMBRAPA responsável pelo projeto – por deixar todos estes registros como "Entre-Safra".

Sugestão: trocar o nome desta variável para ser "Época de Amostragem".

- **Plantação:** são as culturas presentes nos locais das coletas no dia das amostragens.
  - Valores: Milho; Milho/Pasto; Milho/Sorgo; Pasto; Pasto/Pasto; Pasto/Soja; Soja; Soja/Milho; Sorgo; Sorgo/Pasto.

Observação: os valores que possuem duas culturas separadas por "/" referem-se aos registros da época da "Entre-Safra". As rotações definidas que fecham o ciclo do sistema ILP são: "Soja/Milho => Milho/Sorgo => Sorgo/Pasto => Pasto/Soja".

Observação: vale ressaltar que nos primeiros anos do experimento (2005, 2006 e 2007) as rotações eram diferentes, não havia sorgo dentre as culturas produzidas. Todavia, isto não teve impacto na base de dados, pois os registros de 2006 são apenas da época "Na Lavoura". A partir de 2008 que o sorgo começou a ser plantado, determinando a rotação definida anteriormente.

Observação: no ano de 2016 a plantação de soja foi perdida por causa da seca. Por causa disso, naquele ano foi plantado feijão no lugar da soja. Porém, foi decidido em conjunto com o agrônomo da EMBRAPA, que na base de dados tais registros – de todas as épocas de amostragem – ainda seriam anotados com o valor "soja". Desta forma, as culturas e as rotações permaneceram as mesmas.

Observação: no ano de 2019 houve uma tentativa de plantar capim braquiária junto com a soja, no entanto, foi decidido juntamente com a EMBRAPA não reportar este fato na base de dados.

Observação: em 2023 a EMBRAPA deixou de produzir sorgo, assim, tem-se que ficou duas rotações de pasto, como segue: "Milho/Sorgo" virou "Milho/Pasto"; "Sorgo/Pasto" passou a ser "Pasto/Pasto"; e os registros das épocas "Na Lavoura" e "Na Colheita" que seria "Sorgo" passaram a ser "Pasto".

- **Pasto:** refere-se às glebas em que tal espécie de planta daninha havia sido coletada. Como o território do sistema ILP está dividido em quatro glebas, tem-se que os possíveis valores que esta variável pode assumir são: 1, 2, 3 e 4.

Observação: a localização geográfica de cada uma das glebas pode ser encontrada na documentação oficial da EMBRAPA de Milho e Sorgo. https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1106344/1/SistemaILP.pdf

Observação: tem-se que nos primeiros anos do experimento a numeração das glebas era diferente da numeração atual. Tal fato foi levado em consideração e ajustado na base de dados, para que todas as informações ficassem de acordo com a numeração vigente.

Sugestão: trocar o nome desta variável para ser "Gleba".

- **N° de Quadros:** tem-se que as coletas das plantas daninhas sempre são feitas dentro de quadros, isto é, os profissionais da EMBRAPA delimitam dentro de cada gleba alguns quadros (quadrados), e dentro desta área é que é feita a amostragem de plantas daninhas. Assim, tem-se que esta variável remete a quantidade de quadros demarcados para fazer a coleta das amostras. Logo, é uma variável numérica.

Observação: para a maioria dos anos do experimento, foram determinados 10 quadros de 1m² de área para realizar as amostragens – mais informações sobre as áreas do quadros serão explicadas no próximo item.

Observação: em alguns anos a quantidade de quadros foi diferente da padrão – exibida anteriormente. Em 2006 foram demarcados 43 quadros; em 2015, 155 quadros; e em algumas coletas de 2016 foram delimitados 34 quadros. Em todos estes casos a área do quadro era de 0,25m².

- **Área do Quadro (m²):** como mencionado anteriormente, cada quadro marcado para fazer a amostragem tinha uma área definida, logo, tal variável diz respeito a este valor. De modo que é uma variável numérica.

Observação: como dito acima, para os casos em que havia 10 quadros delimitados, a área destes era de 1m² cada. Já para os anos em que a quantidade de quadros foi diferente, a área destes era de 0,25m² cada.

- **Área Total (m²):** refere-se à soma das áreas de todos os quadros utilizados para fazer a coleta dos dados, ou seja, é a área total amostrada. Portanto, esta também é uma variável numérica.

Observação: nos casos de haver 10 quadros de 1m² cada, tem-se que a área total amostrada é de 10m². Já para os outros registros em que a quantidade de quadros era diferente, a área total varia de acordo com as outras duas variáveis dependentes.

- **Observações Gerais:**
  - Em 2006 antes do experimento começar havia Capim Colonião na área. De 2007 em diante foi plantado Capim Mombaça, porém, este começou a aparecer como uma planta daninha em outras áreas. Contudo, como ambos

referem-se à mesma espécie, foi optado por manter todos os registros como Capim Mombaça na base.

1.2. Base de dados do solo

- **Fonte dos dados**: EMBRAPA Milho e Sorgo
- **Formato**: planilhas EXCEL (.xlsx)
- **Total de registros:** 216
- **Anos de coleta**: 2005, 2006, 2012, 2014, 2015, 2016, 2017, 2018, 2019, 2020 e 2022.
- **Variáveis**:
    - **Ano:** refere-se a data em que a amostra de solo foi coletada e analisada. Como as amostras de solo são feitas apenas uma vez por ano, não há anotações sobre o dia e o mês. Os possíveis valores desta variável são os mencionados no item "Período".

Observação: as amostras de solo são feitas apenas uma vez por ano, pois a aplicação dos herbicidas também é feita apenas anualmente.

- **Pasto:** é a mesma variável da base das plantas daninhas, a qual refere-se a gleba em que tal amostra de solo foi coletada. Novamente, seus possíveis valores são 1, 2, 3 e 4.
- **Profundidade (cm):** concerne ao ponto de extração da amostra. Os valores dessa variável estão definidos em intervalos (numéricos), de modo que as opções são: 0-5; 0-10; 0-20; 5-10; 10-20; 15-20; 20-40; 25-30; 40-60 e 45-50.

Observação: a maioria das plantas daninhas encontra-se a uma profundidade de 0-10cm, todavia, optou-se por manter os registros de todas as profundidades para tentar investigar se algumas espécies não morrem com o herbicida por causa de suas raízes serem mais profundas.

- **pH (H2O):** refere-se ao pH da água do solo, por consequência é uma variável numérica com valores entre 0 e 14.

- **H+Al (cmolc/dm3):** variável numérica que quantifica os níveis de hidrogênio com alumínio no solo.
- **Al (cmolc/dm3):** variável numérica que quantifica os níveis de alumínio no solo.
- **Ca (cmolc/dm3):** variável numérica que quantifica os níveis de cálcio no solo.
- **Mg (cmolc/dm3):** variável numérica que quantifica os níveis de magnésio no solo.
- **K (mg/dm3):** variável numérica que quantifica os níveis de potássio no solo.
- **P (mg/dm3):** variável numérica que quantifica os níveis de fósforo no solo.
- **MO (dag/kh):** variável numérica que quantifica os níveis de matéria orgânica no solo.
- **SB (cmolc/dm3):** concerne a soma de bases de cátions. É uma variável numérica.
- **CTC (cmolc/dm3):** é a capacidade de troca de cátions. Também é uma variável numérica.
- **V (%):** denominada saturação por bases, é dada pela razão entre a soma de bases de cátions (SB) e a capacidade de troca de cátions (CTC):

$$V\,(\%) \;=\; \frac{SB}{CTC} * 100$$

  Por consequência, também é uma variável numérica.
- **Sat. Al (%):** refere-se a saturação de alumínio no solo. É uma variável numérica.
- **B (mg/dm3):** variável numérica que quantifica os níveis de boro no solo.
- **Zn (mg/dm3):** variável numérica que quantifica os níveis de zinco no solo.
- **Cu (mg/dm3):** variável numérica que quantifica os níveis de cobre no solo.

- **Mn (mg/dm3):** variável numérica que quantifica os níveis de manganês no solo.
- **Fe (mg/dm3):** variável numérica que quantifica os níveis de ferro no solo.
- **Ca/Mg (cmolc/dm3):** é a razão entre os valores de cálcio e magnésio. É uma variável numérica, por consequência.
- **Ca/K (cmolc/mg):** razão entre os valores de cálcio e potássio. É uma variável numérica, por consequência.
- **Mg/K (cmolc/mg):** razão entre os valores de magnésio e potássio. É uma variável numérica, por consequência.
- **(Ca + Mg)/K (cmolc/mg):** razão entre os valores de cálcio mais magnésio pelos valores de potássio. É uma variável numérica, por consequência.

Observação: para alguns anos não foram calculadas as relações entre os elementos Ca, Mg e K (descritas nos últimos quatros itens da lista), de modo que tais valores foram calculados – via EXCEL – para que todos os registros ficassem completos.

- **Observações Gerais:**
  - Em algumas variáveis – MO, Zn, Cu, Mn, Fe e B – havia dados faltando. Para lidar com esta situação, foram calculadas as medianas de cada uma destas variáveis, e tais valores foram preenchidos nos registros vazios.

1.3. Base de dados do clima

- **Fonte dos dados**: EMBRAPA Milho e Sorgo e Instituto Nacional de Meteorologia (INMET)

Observação: os dados provenientes da EMBRAPA estão divididos em duas planilhas diferentes: a primeira contém informações de uma estação convencional e seus dados referem-se ao período de 01/01/2000 até 18/07/2016; já a segunda, é de uma estação automática, cujos dados são de 11/06/2016 até 31/12/2016. Quanto aos dados do INMET, tem-se que estes remetem aos anos de 2017, 2018, 2019, 2020 ,2021, 2022 e 2023; sendo que cada ano tem sua própria planilha e todas estas são provenientes de uma estação automática.

Logo, percebe-se que foram utilizadas as informações de ambas as fontes, justamente para ter dados climáticos de todo o período do experimento das plantas daninhas para poder realizar o objetivo do projeto.

Observação: as variáveis medidas na estação convencional (EMBRAPA) eram: "Ano"; "Mês"; "Dia"; "Pressão 12h"; "Pressão 18h"; "Pressão 24h"; "B úmido 12h"; "B úmido 18h"; "B úmido 24h"; "Tmax" (Temperatura Máxima); "Tmin" (Temperatura Mínima); "Temp Ar 12h" (Temperatura às 12h); "Temp Ar 18h" (Temperatura às 18h); "Temp Ar 24h" (Temperatura às 24h); "UR 12h" (Umidade Relativa às 12h); "UR 18h" (Umidade Relativa às 18h); "UR 24h" (Umidade Relativa às 24h); "Dir 12h" (Direção do Vento às 12h); "Dir 18h" (Direção do Vento às 18h); "Dir 24h" (Direção do Vento às 24h); "Vento 12h" (Velocidade do Vento às 12h); "Vento 18h" (Velocidade do Vento às 18h); "Vento 24h" (Velocidade do Vento às 24h); "Precipitação"; "Insolação"; e "Evaporação".

Observação: algumas observações devem ser feitas sobre as variáveis da estação convencional:

- Apesar das horas das medições estarem descritas como 12h, 18h e 24h, tem-se que estas foram apenas definidas por convenção, mas que os horários reais das medições eram, na verdade: 9h, 15h e 21h; sendo que a medição das 21h é uma estimativa dos valores do dia todo, com base nas duas medições anteriores.
- "B úmido" refere-se a temperatura de bulbo úmido.
- A direção do vento, neste caso, foi medida em pontos cardeais.

Observação: já as variáveis da estação automática – tanto da EMBRAPA quanto do INMET – são: "Data"; "Hora"; "Tinst" (Temperatura Instantânea); "Tmax" (Temperatura Máxima); "Tmin" (Temperatura Mínima); "URin" (Umidade Relativa Instantânea); "URmin" (Umidade Relativa Mínima); "URmax" (Umidade Relativa Máxima); "Prin" (Pressão Instantânea); "Prmax" (Pressão Máxima); "Prmin" (Pressão Mínima); "Vvel" (Velocidade do Vento); "Vdir" (Direção do Vento); "Vraj" (Rajada de Vento); "Rad" (Radiação); "Chuva"; "Ovrin" (Ponto de Orvalho Instantâneo); "Ovrmax" (Ponto de Orvalho Máximo); e "Orvmin" (Ponto de Orvalho Mínimo).

Observação: tem-se algumas considerações sobre estas variáveis:

- Nesta estação a direção do vento era medida em graus.

Observação: para elaborar a base de dados do clima, foi feita a junção das bases da EMBRAPA com as bases do INMET. Para isso, foram selecionadas as variáveis comuns a ambas. Abaixo segue o processo de conjunção de cada uma das variáveis da estação convencional para se equiparar com as variáveis da estação automática ou vice-versa:

- Os valores da "Data" foram desmembrados para que as informações "Dia", "Mês" e "Ano" ficassem separadas.
- A média dos valores de "Pressão 12h", "Pressão 18" e "Pressão 24h" foi calculada e definida como "Pressão Instantânea".
- A média de "Temp Ar 12h", "Temp Ar 18h" e "Temp Ar 24h" foi calculada e definida como "Temperatura Instantânea".
- A média de "UR 12h", "UR 18h" e "UR 24h" foi calculada e definida como "Umidade Relativa Instantânea".
- Foi feita uma tratativa na planilha da estação convencional, de modo que os pontos cardeais foram convertidos para graus (N = 360°, NO = 315°, O = 270°, SO = 225°, S = 180°, SE = 135°, L = 90°, NE = 45°). A partir disso a média de "Dir 12h", "Dir 18h" e "Dir 24h" foi calculada e definida como "Direção do Vento".
- A média de "Vento 12h", "Vento 18h" e "Vento 24h" foi calculada e definida como "Velocidade do Vento".
- Não foi feita a conversão de "Insolação" para "Radiação", pois a insolação era medida por um heliógrafo, ou seja, a luz emitida queimava uma fita durante um certo intervalo de tempo para medir tal variável; enquanto a radiação referia-se apenas ao momento da medição. Como as variáveis não estavam sendo medidas na mesma proporção de tempo, essa conversão não foi realizada.
- Já as variáveis "Precipitação", "Tmax" e "Tmin" puderam ser juntadas diretamente com seus correspondentes da estação automática.
- As demais variáveis de ambas as fontes – "B úmido 12h", "B úmido 18h", "B úmido 24h", "Evaporação", "Hora", "URmin", "UR max", "Prmax", "Prmin", "Vraj", "Orvin", "Orvmax", e "Orvmin" – foram descartadas por não haver dados equivalentes na outra base para fazer a junção.

- **Formato**: planilhas EXCEL (.xlsx)
- **Total de registros:** 8067
- **Período**: 2000 até 2023
- **Variáveis:**
    - **Ano:** refere-se ao ano em que foram feitas as medições das variáveis de clima. Seus possíveis valores são todos os anos de 2000 até 2023 – valores advindos da junção de todas as bases.
    - **Mês:** refere-se ao mês em que foram feitas as medições das variáveis de clima. Tal variável foi anotada como numérica, de modo que seus valores vão de 1 a 12.
    - **Dia:** refere-se ao dia em que foram feitas as mediçẽs das variáveis de clima. Tal variável foi anotada como numérica, de modo que seus valores vão de 1 até 31.
    - **Pressão (hPa):** é o valor da pressão registrado no momento da medição, ou seja, é a pressão instantânea. É uma variável numérica.
    - **Temperatura Média (°C):** é o valor da temperatura registrado no momento da medição, ou seja, é a temperatura instantânea. É uma variável numérica.
    - **Temperatura Máxima (°C):** é a temperatura mais alta registrada durante o dia. É uma variável numérica.
    - **Temperatura Mínima (°C):** é a temperatura mais baixa registrada durante o dia. É uma variável numérica.
    - **Umidade Relativa (%):** é o valor da umidade do ar registrado no momento da medição, ou seja, é a umidade relativa instantânea. É uma variável numérica.
    - **Velocidade do Vento (m/s):** é o valor da velocidade do vento no momento da medição. É uma variável numérica.
    - **Direção do Vento (°):** é o valor da direção do vento no momento da medição. É uma variável numérica.

- **Precipitação (mm):** é a quantidade de chuva que ocorreu durante o dia. Também é uma variável numérica.

- **Observações Gerais:**

  - Em todas as planilhas – tanto da estação convencional, quanto da estação automática – havia dados faltando. Todavia, a tratativa para estes registros foi feita apenas depois da junção das bases. Isto é, primeiro as planilhas foram compiladas e depois foram analisados quais registros ainda estavam faltando na base de clima unificada.

  - Para estes valores que estavam faltando, foram calculadas as medianas das respectivas variáveis e tais valores foram preenchidos nestes registros em branco.

1.4. Base de dados unificada

- **Formato**: planilha EXCEL (.xlsx)

- **Total de registros:** 1541

- **Anos de coleta**: 2006, 2015, 2016, 2017, 2018, 2019, 2020 e 2022

- **Variáveis:** as variáveis da base unificada são todas as variáveis presentes nas bases das plantas daninhas, do solo e do clima; as quais já foram explicadas em mais detalhes anteriormente.

Observação: a base unificada conta tanto com a variável "Data", quanto com a variável "Ano". Isso porque as bases do clima e das plantas daninhas continham a data completa, enquanto a base do solo tinha apenas o ano. Assim, por causa do processo de junção das bases, ambas as informações aparecem na base unificada.

- **Observações Gerais:** abaixo segue o processo feito nas bases de dados para elaboração da base unificada:

  - Primeiro, foi necessário agrupar os registros da base das plantas daninhas. Para isso, os registros foram agrupados por pasto, por data e por espécie. Dessa forma, os registros de mesmas espécies – de mesmo pasto e mesma data

– foram somados. Por exemplo: antes havia três registros (separados) da espécie caruru no dia 12/03/2015 no pasto 2; com o agrupamento ficou apenas um registro dizendo que no dia 12/03/2015 no pasto 2 havia três carurus, ou seja, as quantidades de plantas daninhas foram somadas.

- Como a variável "Data" é uma informação em comum das bases das plantas daninhas e do clima, foi possível fazer uma junção direta entre estas bases. Isto é, para uma mesma data de coleta de plantas daninhas, as informações do clima desse mesmo dia foram concatenadas com as informações de plantio. Assim, as colunas da base do clima foram adicionadas na base das plantas daninhas.

- Em paralelo, os registros da base do solo também foram agrupados para facilitar a junção. Os registros foram agrupados por pasto, por ano e por profundidade. Assim, registros de uma mesma profundidade – de um mesmo pasto e mesmo ano – tiveram as médias de seus valores calculados, para ficar um só registro.

- Depois disso, foi a junção "de todos para todos" da base do solo com a planilha que continha as informações de climas e das plantas daninhas. Essa junção foi feita a partir do ano, que era a informação comum a ambos os conjuntos de dados.

- Como há vários intervalos de profundidades, tem-se que cada registro das plantas daninhas foi repetido para as diferentes profundidades. Isso foi feito para considerar que os valores de solo encontrados nas distintas profundidades ainda referem-se à mesma planta daninha encontrada no solo. Além disso, tal processo também contribuiu para aumentar a quantidade de dados da base unificada.

Observação: vale ressaltar que para fazer a leitura das bases, para rodar os algoritmos de predição, foi feito um *encoding* dos dados; isto é, os valores nominais eram transformados em numéricos para ficarem em um formato mais adequado para o computador.

Sugestão: todas as bases estão no formato de planilhas, porém, poderia ser feito um banco de dados relacional (SQL) para facilitar a integração da base com uma plataforma de coleta de dados, de modo a organizar melhor todas as informações.

2. GUIA DE BOAS PRÁTICAS PARA COLETA DE DADOS

Durante o desenvolvimento do trabalho – em particular durante a elaboração das bases de dados –, percebeu-se alguns problemas nos processos de coleta dos dados. Tais situações tornaram-se obstáculos que tiveram de ser superados durante o desenvolvimento do projeto e também contribuíram negativamente para o desempenho dos algoritmos de aprendizado de máquina desenvolvidos.

Assim, esta seção apresenta um guia de boas práticas a serem respeitadas durante a coleta das amostras. Tais diretrizes foram baseadas em alguns aspectos observados no processo de coleta de dados atual, e que podem ser melhorados para aprimorar os resultados deste projeto e contribuir numa melhor performance de trabalhos futuros.

- **Padronizar a formatação:** é importante que as planilhas nas quais serão anotados os dados coletados tenham sempre o mesmo formato, isto é, que as colunas sejam sempre escritas na mesma ordem. De modo que na hora de fazer a leitura desse documento seja fácil de localizar as informações desejadas. Tal ponto é importante, pois quando os documentos não dispõem as variáveis na mesma ordem, leva-se mais tempo para encontrar os dados procurados; além de que pode fazer com que alguma informação relevante seja perdida.

Sugestão: criar um modelo base com todas as variáveis na ordem que serão coletadas, e sempre utilizar esse modelo de documento na hora das coletas; não criar arquivos novos com formatações diferentes para cada coleta. Garantir que este modelo disponha os dados de forma organizada e fácil de compreender.

- **Padronizar as variáveis:** além da padronização da formatação dos documentos, é importante que as informações a serem anotadas, também sejam as mesmas – sempre que possível. Documentos que não apresentam as mesmas variáveis podem gerar disparidades na base de dados, ou podem fazer com que tais informações tenham de ser descartadas. Por exemplo, se queremos saber em qual época de amostragem tal coleta foi realizada, mas tal informação não aparece em certas planilhas, os dados destes documentos terão que ser desconsiderados pois não apresentam uma informação vital para o projeto. Ou poderia-se tentar deduzir em que época de amostragem estes dados referem-se, porém, isto pode gerar informações falsas, que podem prejudicar o trabalho ainda mais; além de também exigir mais tempo e

retrabalho.

Sugestão: o ideal para evitar esse problema também seria criar um modelo base com todas as variáveis que devem ser coletadas, e sempre utilizar esse modelo para garantir que todas as informações necessárias sejam coletadas.

Sugestão: evitar anotar informações que não serão utilizadas no desenvolvimento do projeto. Ao menos que estas revelem-se importantes para o projeto, neste caso, reavaliar os modelos padrões utilizados e os dados previamente coletados.

Observação: caso não seja possível coletar algum dado, é importante definir algum valor/símbolo como nulo, para evitar deixar espaços em branco nas planilhas. Quanto maior a quantidade de dados disponíveis para análise melhor, pois evita que tratativas tenham que ser feitas para lidar com estes valores faltantes.

- **Padronizar a nomenclatura:** com as variáveis bem definidas, também é importante que os possíveis valores destas também estejam bem determinados. Por exemplo, tem-se que mesmas espécies de plantas daninhas apresentam nomes comuns diferentes. Entretanto, se em um registro está escrito “capim colonião” e em outro está escrito “capim mombaça”, o computador ao transformar estas informações em valores numéricos, irá atribuir a cada um destes dados ID 's diferentes. Isto porque a máquina não sabe que estas plantas são da mesma espécie. Ou até mesmo um profissional que não é da área poderia confundir e tratar estas informações como coisas diferentes. Assim, é importante sempre definir quais os possíveis valores que uma variável pode assumir e sempre usar estes mesmos valores.

Observações: esta prática é mais pertinente quando estamos tratando de dados nominais.

Sugestão: fazer uma lista com os possíveis valores que um determinada variável pode assumir e repassar para todos os profissionais envolvidos no projeto, para que sempre usem tais nomenclaturas. No caso de uma variável poder assumir um valor que ainda não apareceu – por exemplo, uma espécie de planta daninha que nunca havia aparecido antes – definir qual será o seu nome e acrescentar na lista.

Sugestão: no caso da base das plantas daninhas é importante padronizar se a nomenclatura das espécies será apenas o seu nome comum ou o seu nome científico, pois tem-se que ambas as opções apareceram nas planilhas disponibilizadas pela EMBRAPA. No caso de optar-se

pelo nome comum, deve-se padronizar qual será este, caso haja mais de uma possibilidade. Por exemplo, se uma planta coletada será registrada apenas como “caruru” ou como “caruru rosa”. Pois na primeira abordagem tem-se que todas as plantas da família “caruru” serão agrupadas em um único grupo, enquanto na segunda os agrupamentos serão por aquela espécie em específico.

- **Padronizar a escrita:** depois de definir quais são os possíveis valores que uma determinada variável pode assumir, é importante também determinar como estes devem ser escritos. Por exemplo, é fácil ver que os termos “joá de capote” e “joá-de-capote” referem-se à mesma planta daninha. Entretanto, por estarem escritos de formas diferentes, na hora em que o computador converter estes registros para numéricos, seus ID 's serão diferentes. Isso porque o computador não faz uma leitura do conteúdo, mas sim uma análise dos caracteres. Logo, como os caracteres de “joá de capote” e “joá-de-capote” não são os mesmos, a máquina interpretará essas informações como coisas distintas. Portanto, a padronização da escrita dos possíveis valores de cada variável também é de extrema importância.

Sugestão: na hora em que criar a lista com os possíveis valores de cada variável, já determinar como deve ser sua escrita e garantir que todos os envolvidos no projeto estejam a par dessas diretrizes.

- **Padronizar as unidades de medida:** no caso de variáveis numéricas, é importante garantir que valores de uma mesma variável sejam sempre anotados numa mesma unidade de medida. Pois, quando isso não ocorre acaba sendo necessário converter os valores para que todos fiquem na mesma unidade. Consumindo tempo e recursos do projeto.

Importante: sempre anotar qual é a unidade de medida que está sendo usada, pois dificulta a utilização e interpretação dos dados quando isso não é feito.

Observação: tem-se que estas práticas são fáceis de serem implementadas em coletas de dados de uma mesma fonte. Contudo, caso seja um projeto que vai envolver dados de mais de uma fonte, é importante tentar ao máximo garantir que as variáveis, nomenclaturas, formas de escritas e unidades, sejam as mesmas em ambas as bases. Caso isso não seja possível, manter registros detalhados de todas essas informações, para facilitar processos de equivalências e conversões.

Contudo, tem-se que o maior problema observado é a escassez de dados. Verifica-se que tanto as bases individuais, bem como a base unificada apresentam uma quantidade insuficiente de dados. Isso torna-se um empecilho para o desenvolvimento do projeto, pois os algoritmos de aprendizado de máquina requerem um grande volume de dados para terem um desempenho satisfatório. Vale destacar que quando tais modelos são desenvolvidos, uma parte do conjunto de dados é destinada para o treinamento do modelo e o restante é utilizado nas fases de teste e/ou validação e teste.

Como exemplo, tem-se que na dissertação desenvolvida a proporção entre os dados utilizados foi de 70% para treinamento e 30% para teste. Isto é, apenas 30% dos 1514 registros da base unificada passaram pelo processo de predição, isto equivale à aproximadamente 462 registros. Tal valor não é o bastante para aferir relações mais profundas entre os dados. Tanto que o objetivo inicial do projeto (e da dissertação) era desenvolver um modelo de predição das espécies de plantas daninhas. Porém, como a quantidade de dados era insuficiente, não foi possível averiguar os fatores ambientais que favoreciam mais uma espećie ou outra, de modo que outros modelos de predição tiveram de ser elaborados.

Assim, faz-se necessário incrementar a(s) base(s) de dados com mais informações. Para isso, é preciso que mais coletas de dados sejam realizadas. No experimento das plantas daninhas, tem-se que as amostragens eram feitas três ou quatros vezes por ano numa área total de 10m². Para atingir uma quantidade satisfatória de dados, tais coletas precisam ser feitas com mais frequência. Além disso, a área total amostrada também pode ser maior, de modo a aumentar as chances de encontrar plantas daninhas para compor os registros da base.

Do mesmo modo, a base de dados do solo também precisaria ser consideravelmente maior. Apesar de haver um motivo para as amostras de solo serem feitas apenas uma vez por ano, tal fato prejudica o desempenho dos algoritmos em diversos aspectos: primeiro, a falta de dados dificulta tirar alguma conclusão sobre a influência dos fatores de solo nas espécies de plantas daninhas.

Outro ponto, é que dessa forma não é possível assegurar que a relação entre os dados seja verdadeira, pois se a amostra de plantas daninhas foi feita em abril e a coleta do solo foi feita em novembro – por exemplo – não há garantias de que as propriedades do solo tenham se mantido as mesmas durante todo esse tempo. Logo, afirmar algum relação entre os fatores de solo e as determinadas espécies de plantas daninhas, pode não ser realista nesses termos;

pois grandes são as chances de que os elementos do solo na data em que a coleta de plantas daninhas foi feita eram outros dos que foram verificados na época da amostragem do solo.

Tem-se que as datas das coletas dos dados de plantas daninhas e do solo não coincidem, porém esta é a única variável em comum das três bases individuais. Decorrente disso, na hora de realizar a junção das bases, muitos dados foram perdidos, pois tais registros de solo e de plantas daninhas não tinham datas em comum para serem concatenados.

Por isso que o processo de agrupamento foi feito: para facilitar o processo de junção dessas bases. Todavia, tem-se que como só há um registro dos dados do solo para cada ano, tem-se que estes foram concatenados com todos os dias daquele mesmo ano em que houveram coletas de plantas daninhas; porém, isto retoma a questão da veracidade das relações – mencionada antes.

Tem-se que o cenário ideal seria que as coletas de plantas daninhas e de solo ocorressem sempre no mesmo dia, de modo a aumentar a quantidade de dados de ambas as bases e, também, facilitar o processo de junção das mesmas. Vale ressaltar que a base do clima não apresenta empecilhos nesse processo, pois – neste caso – tem-se os registros de todos os dias do ano.

3. CONSIDERAÇÕES FINAIS

Este documento apresentou um relatório sobre como as bases de dados – desenvolvidas durante o projeto realizado pela UNIVALI em parceria com a EMBRAPA – estão estruturadas, bem como todas as tratativas realizadas em cada um destes conjuntos de dados para poder elaborar a base de dados unificada.

Além disso, também foi produzido um guia de boas práticas para coleta de dados, para ajudar nesse processo; melhorar o desempenho dos algoritmos desenvolvidos durante o trabalho; e possibilitar o desempenho de novos modelos de aprendizado de máquina com outros objetivos.

Os modelos de aprendizado de máquina de predição de cultura e de época de amostragem, elaborados durante o desenvolvimento do projeto, estão disponíveis em: https://github.com/anagomes04/projeto-embrapa-dissertacao. As bases de dados não estão disponíveis publicamente, devido aos direitos autorais dos dados.