capítulo 5

Perspectivas dos agrodados e da digitalização da agricultura

Jayme Garcia Arnal Barbedo
Luciana Alvim Santos Romani
Debora Pignatari Drucker
Carla Geovana do Nascimento Macário

Introdução

Com a crescente demanda por alimentos e por uma produção social e ambientalmente mais sustentável, cresce também a demanda por soluções tecnológicas capazes de resolver os muitos desafios envolvidos nessa empreitada. Com o grande avanço observado na inteligência artificial na última década, tornou-se viável vencer muitos desses desafios. Porém, para que modelos de inteligência artificial funcionem adequadamente, eles precisam ser treinados com dados confiáveis de boa qualidade que representem adequadamente o problema que se deseja resolver. Esse é certamente o maior desafio para tornar tais tecnologias uma realidade, não apenas porque a quantidade de dados necessária normalmente é muito grande, mas também porque o ambiente agrícola é frequentemente de difícil acesso e sujeito a intempéries e a outros fenômenos ambientais adversos.

Este capítulo tem como objetivo explorar as perspectivas para a coleta e a utilização dos dados agrícolas, sendo dividido em duas partes. A primeira parte explora como será realizada a governança dos dados no âmbito do Centro de Ciência para o Desenvolvimento em Agricultura Digital (CCD-AD/Semear Digital). Esse Centro é uma iniciativa financiada pela Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp), cujo objetivo principal é atuar em pesquisa, desenvolvimento e inovação em tecnologias emergentes visando, principalmente, à inclusão digital de pequenos e médios produtores rurais de modo a ter ganhos de produtividade e de competitividade, maior impacto econômico em termos de aumento de escala, redução de custos e melhoria na eficiência da produção agrícola nos diversos elos das cadeias produtivas, de maneira sustentável, tanto do ponto de vista econômico, quanto ambiental e social. O CCD-AD/Semear Digital desenvolverá suas atividades em dez áreas distribuídas por todas as regiões do Brasil, denominadas Distritos Agrotecnológicos (DATs), e para que a coleta, a gestão e a utilização dos dados sejam realizadas de maneira adequada, protocolos precisarão ser definidos e implantados. A segunda é dedicada aos desafios, potenciais tendências e oportunidades relacionados aos dados agrícolas, abordando questões como sensores utilizados para coleta dos dados, técnicas para extração de informação relevante, barreiras para a obtenção dos dados e possíveis soluções para os problemas que ainda dificultam a digitalização no campo.

Governança de dados agrícolas no âmbito do Centro de Ciência para o desenvolvimento em agricultura digital

A Fapesp foi a primeira agência de fomento à pesquisa nacional a estabelecer diretrizes quanto ao adequado tratamento dos dados gerados em projetos por ela financiados. A exigência de planos de gestão de dados acompanhando as submissões de propostas passou a ser feita em todos os programas e linhas de financiamento, da mesma forma que é praticada pelas principais agências de fomento internacionais, desde o início dos anos 2010. Em linhas gerais, planos de gestão de dados descrevem os dados que serão gerados pelo projeto, seu volume, formato, estrutura, como serão obtidos ou produzidos. Também descrevem quais as estratégias para o armazenamento e a preservação dos dados ao longo do tempo, regras e restrições para controle e acesso, dentre outros aspectos, sendo ferramentas valiosas para orientar a governança de dados. Há orientações mais específicas para propostas de Centros, como é o caso do CCD-AD/Semear Digital, os quais devem seguir as "Diretrizes para Planos de Gestão de Dados para Propostas de Centros da Fapesp" (Fundação de Amparo à Pesquisa do Estado de São Paulo, 2021).

A exigência de elaboração desse documento auxiliou a equipe envolvida com a proposta a prever, desde o início, algumas questões-chave que orientam a governança dos dados do Centro. Primeiramente, foi preciso descrever resumidamente os dados a serem produzidos e gerenciados pelo CCD-AD. A Agricultura Digital é multi e interdisciplinar e, consequentemente, o CCD-AD gera dados de diferentes áreas relevantes para desenvolvimento agrícola, tais como clima, geotecnologias, mercado, distribuição e logística. Tais dados são utilizados para gerar conhecimentos aplicados em todos os elos da cadeia produtiva, desde a pré-produção, passando pela produção até a fase de pós-produção. São gerados dados numéricos, textuais, imagens coloridas RGB e multiespectrais, muitos deles georreferenciados, a partir de diferentes dispositivos, como sensores meteorológicos e ambientais, câmeras, aeronaves e robôs, sensores remotos e celulares. Há também a compilação de dados preexistentes, em especial nas áreas dos DATs. Será priorizado o uso de formatos de dados abertos e interoperáveis, seguindo os princípios FAIR (*Findable, Accessible, Interoperable, Reusable*) (Wilkinson et al., 2016), incluindo os casos de restrições comerciais e da Lei Geral de Proteção de Dados Pessoais (LGPD).

Em relação à gestão dos dados, eles serão preservados nos repositórios institucionais da Embrapa e de parceiros, acompanhados de metadados que os

descrevem. Os repositórios institucionais implementam padrões de dados e de metadados de acordo com o domínio de conhecimento e contam com mecanismos de restrição de acesso sempre que necessário. Também contam com mecanismos de backup regulares e uma cópia adicional dos dados será guardada com o coordenador do projeto. Os scripts analíticos em linguagem de programação R, Python e outros, bem como softwares gerados no contexto do Centro, serão também preservados e documentados de acordo com as melhores práticas preconizadas globalmente. Na área de aprendizado de máquina, é adotada a proposta de Gebru et al. (2018), respondendo a questões como "por que o conjunto de dados foi criado?", "qual pré-processamento foi feito?" ou "o conjunto de dados será atualizado, e com qual frequência?".

O Redape – Repositório de Dados de Pesquisa da Embrapa[1], implementa grande parte dos princípios FAIR, como a atribuição de um identificador globalmente único e persistente aos conjuntos de dados; o registro e indexação de dados e metadados em um recurso pesquisável; os dados e metadados são acompanhados de uma licença de uso de dados clara e acessível; e os dados e metadados atendem aos padrões da comunidade que são relevantes para cada domínio, viabilizando o seu reuso. Considerando que os produtos de dados gerados ou adquiridos pelo Centro deverão ser diversos, podendo abranger levantamentos de campo (nato-digitais ou não), modelos, algoritmos, gráficos, mapas, vídeos, planilhas, gravações de áudio, entre outros, em que diferentes padrões de metadados e vocabulários controlados poderão ser utilizados para representar as informações geradas.

O compartilhamento dos dados respeita, quando aplicável, a Política de Dados, Informação e Conhecimento da Embrapa (Embrapa, 2019), que estabelece diretrizes para restrição de acesso por motivos de propriedade intelectual, privacidade de dados e outros previstos por lei, bem como período de embargo para projetos em andamento. Os conjuntos de dados publicados são acompanhados por licenças de uso, de acordo com as diretrizes institucionais, e seu impacto na comunidade será monitorado de acordo com os indicadores estabelecidos institucionalmente. As regulamentações das instituições parceiras são também respeitadas. Em relação às restrições éticas, de confidencialidade e legais, dados gerados em parceria com instituições privadas podem ser protegidos. Dados que envolvam entrevistas com pessoas ou levantamentos de propriedades rurais são obtidos sob termos de consentimento e de acordo com a LGPD. Dados obtidos por meio de aplicativos

[1] Disponível em: www.embrapa.br/redape.

móveis que envolvam informações pessoais e confidenciais, como, por exemplo, transações financeiras, não serão passíveis de acesso, exceto no caso de dados anonimizados, isto é, agregados espacial e temporalmente.

Esse conjunto de definições auxiliam a governança dos dados de forma a assegurar sua integração, uso e reúso para atender de forma confiável e transparente aos desafios do CCD-AD e da sociedade como um todo.

Tendências, desafios e oportunidades

A revolução que vem sendo observada com a evolução das tecnologias baseadas em inteligência artificial é, em grande parte, devida à enorme quantidade de dados que vem sendo gerada em diferentes contextos e circunstâncias. A internet, em particular, tornou possível a coleta de dados em quantidades nunca antes vistas, especialmente pelo uso de ferramentas de buscas e redes sociais (Knoke; Yang, 2008). Com o advento dos algoritmos de aprendizado profundo no início da década de 2010, a análise de dados e extração de informação se tornou muito mais eficiente, levando ao surgimento de tecnologias com um amplo espectro de aplicações (Deng; Yu, 2014). No caso da agricultura, embora se fale muito no potencial da inteligência artificial e haja muita atividade acadêmica no assunto, o número de tecnologias deste tipo sendo utilizadas na prática ainda é relativamente pequeno (Barbedo, 2022b).

O que ocorre é que o ambiente agrícola possui peculiaridades normalmente não encontradas em outras circunstâncias. Em primeiro lugar, o campo é um ambiente não-controlado e não-estruturado, fazendo com que haja uma variedade muito grande de condições relacionadas a fatores como iluminação, clima, características do solo, cultivares, estágio de desenvolvimento das plantas, relevo, entre muitas outras. Para que uma base de dados represente adequadamente um determinado problema, todas essas condições precisam estar contempladas. Isto é particularmente verdadeiro no caso de imagens, porque em geral os modelos de inteligência artificial não são muito acurados ao analisar imagens com características diferentes daquelas utilizadas no seu treinamento (Barbedo, 2018). Como resultado, bases realmente representativas devem conter uma quantidade elevada de dados confiáveis coletados em uma grande variedade de locais e condições. Este já seria um desafio em ambientes controlados e estruturados, mas se torna particularmente complexo em um ambiente que frequentemente é de difícil acesso, com conectividade limitada, sujeito a intempéries e com terreno acidentado (Barbedo, 2022b). A maior parte das

tendências relacionadas a dados agrícolas observadas atualmente visa superar algumas das barreiras ou mitigar seus efeitos. A seguir são apresentadas algumas das principais tendências observadas atualmente:

- A quantidade de sensores móveis e fixos no campo vem aumentando continuamente, e essa tendência deve continuar. Há vários fatores que contribuem para isso, incluindo o crescente número de produtores dispostos a utilizar essas tecnologias, miniaturização e barateamento dos equipamentos, evolução das tecnologias de conectividade para transmissão dos dados, surgimento de sistemas de proteção a intempéries mais eficientes, entre outros. Além disso, a coleta de imagens se tornou muito mais viável com a popularização de celulares equipados com câmeras digitais de boa qualidade, bem como com o aumento no uso de veículos aéreos não tripulados, também conhecidos como drones.
- Sensores embarcados em maquinário agrícola vêm se tornando cada vez mais comuns, sendo esta uma fonte valiosa de diferentes tipos de dados. Sensores embarcados têm a vantagem de coletar dados de maneira autônoma e sem afetar a operação normal da fazenda, e como máquinas agrícolas normalmente percorrem toda a propriedade poucas áreas ficam sem cobertura. Além disso, sensores acoplados a modelos de inteligência artificial e atuadores já vêm sendo usados, por exemplo, para detectar ervas daninhas e eliminá-las em tempo real, sem qualquer intervenção humana (Wu et al., 2021).
- Em muitos casos, é possível processar os dados à medida que eles vão sendo coletados, sendo armazenada somente a informação útil a ser usada na tomada de decisões. Esse tipo de abordagem, chamado de computação de borda (Satyanarayanan, 2017), vem sendo utilizada cada vez com mais frequência por reduzir substancialmente a quantidade de dados a ser armazenada, além de exigir níveis de conectividade muito mais básicos e baratos. Este tipo de abordagem se tornou viável após o surgimento de computadores miniaturizados de custo baixo e com alta capacidade de processamento, como o Raspberry Pi e o BeagleBoard.
- Alguns problemas agrícolas são tão complexos que uma única fonte de dados pode não ser suficiente para a geração de modelos ou tecnologias eficazes, mesmo que os dados cubram toda a variabilidade associada ao problema. Nesses casos, pode ser necessário combinar diferentes tipos de dados em uma única solução empregando uma abordagem mais sistêmica. Esse tipo de estratégia, chamada fusão de dados, vem sendo utilizada há bastante tempo no contexto do sensoriamento remoto, mas vem também rapidamente ganhando espaço em outros contextos agrícolas (Barbedo,

2022a). Existem diversas técnicas específicas para esse fim, sendo que a melhor abordagem depende da natureza dos dados que se pretende combinar. Dentre os dados que vêm sendo combinados, pode-se destacar diferentes imagens com dados meteorológicos, bem como diferentes tipos de imagens.

- Como comentado anteriormente, técnicas de aprendizado profundo vêm sendo largamente utilizadas em uma variedade de aplicações, tanto na agricultura quanto em outros setores. À medida que as arquiteturas de redes profundas evoluem, essa tendência tende a se intensificar ainda mais, sendo particularmente forte no caso de imagens digitais, uma vez que os modelos de aprendizado profundo extraem informação relevante diretamente a partir das imagens, sem a necessidade de se extrair parâmetros e atributos específicos para cada fim (Barbedo, 2022b).

Para que essas tendências levem mesmo a avanços significativos, alguns desafios precisam ser vencidos. Modelos computacionais em geral necessitam de uma quantidade considerável de dados para funcionar corretamente. Quanto maior a variabilidade associada ao problema, mais amostras confiáveis são necessárias. Isso é particularmente ruim no caso de aplicações agrícolas, porque o número de fatores que introduzem variabilidade é muito grande (Barbedo, 2018). Não há soluções simples para o problema, mas algumas alternativas vêm sendo aplicadas. O compartilhamento de dados vem sendo amplamente incentivado por agências de fomento e periódicos científicos, o que pode não apenas aumentar a quantidade de dados disponível, mas também aumentar a variabilidade desses dados uma vez que provavelmente foram capturados em diferentes localizações geográficas. É importante, porém, que tais dados sejam compartilhados seguindo os princípios FAIR atendendo a padrões de encontrabilidade, acessibilidade, interoperabilidade e reusabilidade.

Outra maneira de aumentar tanto a qualidade quanto a variedade dos dados é envolver indivíduos fora da comunidade científica nos esforços de construção das bases de dados, usando os princípios da ciência cidadã (Irwin, 2002; Silvertown, 2009) ou *crowdsourcing*. Há muitos incentivos que podem ser aplicados a fim de engajar as pessoas, incluindo mecanismos de recompensas extensivamente usados em redes sociais.

Alguns sensores que vêm sendo aplicados na agricultura geram uma quantidade muito grande de dados. Câmeras hiperespectrais são um bom exemplo, já que uma única imagem pode ter perto de 1 Gigabyte (GB). Uma solução já mencionada é

a utilização de computação de borda, em que o processamento é realizado em tempo real sem a necessidade de armazenamento de longo prazo e de transmissão de todos os dados. Porém, há situações em que uma infraestrutura robusta de armazenamento é inevitável, com custos bastante elevados. Apesar da capacidade de armazenamento continuar crescendo e os preços diminuindo, a quantidade de dados gerada vem também aumentando, talvez a um ritmo ainda maior, o que é um fator a ser considerado na aplicação das tecnologias na agricultura.

Portanto, o problema dos dados na agricultura ainda está longe de ser resolvido. Embora essa situação não seja ideal, ela traz diversas oportunidades tanto para cientistas quanto para empresas interessadas em desenvolver soluções que realmente atendam às necessidades dos produtores, numa área de estudo ainda não consolidada e com concorrência baixa. É importante o desenvolvimento de novas tecnologias que considerem diferentes demandas de todos os produtores rurais, uma vez que a inserção de pequenas e médias propriedades não apenas promove a inclusão desses proprietários, como também promove ganhos de escala, tornando a tecnologia potencialmente muito mais rentável.

Conclusões

A crescente disponibilidade de dados levou ao surgimento de uma gama de tecnologias e de aplicações que vêm tendo um grande impacto em todos os setores da sociedade. O número relativamente baixo de tecnologias baseadas em inteligência artificial sendo efetivamente aplicadas na prática na agricultura não é devido a uma falta de interesse ou de esforços para seu desenvolvimento, mas sim à dificuldade de se conseguir dados em quantidade e qualidade suficientes para que tais tecnologias se tornem uma realidade. Iniciativas como o CCD-AD/Semear Digital têm como uma de suas premissas básicas o estabelecimento de condições apropriadas para a obtenção, a gestão e a utilização de dados agrícolas que sejam de fato úteis para a criação de novas tecnologias. À medida que a tecnologia dos sensores e os métodos para processamento dos dados evoluem, esses objetivos se tornam mais factíveis. Porém, tais esforços dependem do estabelecimento de protocolos e de diretrizes que guiem a obtenção e governança dos dados, sempre com foco em dados de qualidade. Este capítulo teve como objetivo descrever os esforços que vêm sendo feitos para garantir uma governança adequada dos dados dentro de um centro de pesquisa real, bem como traçar um panorama para os dados

agrícolas considerando a situação atual e as tendências que vêm sendo observadas ao longo dos últimos anos. Apesar das dificuldades impostas pelo ambiente agrícola, o uso crescente de dados e tecnologias relacionadas é algo irreversível. Esse cenário traz um grande potencial para melhoria da qualidade de vida e da rentabilidade dos produtores rurais, porém será necessário um esforço, tanto por parte dos donos das tecnologias quanto pelo governo, para que tais benefícios atinjam a todos.

Referências

BARBEDO, J. G. A. Factors influencing the use of deep learning for plant disease recognition. **Biosystems Engineering**, v. 172, p. 84-91, Aug. 2018. DOI: https://doi.org/10.1016/j.biosystemseng.2018.05.013.

BARBEDO, J. G. A. Data fusion in agriculture: resolving ambiguities and closing data gaps. **Sensors**, v. 22, n. 6, 2285, 2022a. DOI: https://doi.org/10.3390/s22062285.

BARBEDO, J. G. A. Deep learning applied to plant pathology: the problem of data representativeness. **Tropical Plant Pathology**, v. 47, n. 1, p. 85-94, 2022b. DOI: https://doi.org/10.1007/s40858-021-00459-9.

DENG, L.; YU, D. Deep learning: methods and applications. **Foundations and Trends in Signal Processing**, v. 7, n. 3-4, p. 197-387, 2014. DOI: https://doi.org/10.1561/2000000039.

EMBRAPA. Política de dados, informação e conhecimento da Embrapa, de 4 de abril de 2019. **Boletim de Comunicações Administrativas**, ano 45, n.16, p. 2-19, abr. 2019. Disponível em: https://www.embrapa.br/politica-de-governanca-de-dados-informacao-e-conhecimento. Acesso em: 5 dez. 2023.

FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO. **Gestão de dados**: diretrizes para Planos de Gestão de Dados (PGD) para propostas de centros. 2021. Disponível em: https://fapesp.br/14974/diretrizes-para-planos-de-gestao-de-dados-pgd-para-propostas-de-centros. Acesso em: 5 dez. 2023.

GEBRU, T.; MORGENSTERN, J.; VECCHIONE, B.; VAUGHAN, J. W.; WALLACH, H.; DAUMÉ III, H.; CRAWFORD, K. Datasheets for datasets. **Communications of the ACM**, v. 64, n. 12, p. 86-92, Dec. 2021. DOI: https://doi.org/10.1145/3458723.

IRWIN, A. **Citizen science:** a study of people, expertise and sustainable development**.** 1. ed. London: Routledge, 2002.

KNOKE, D.; YANG, S. **Social network analysis**. 2. ed. Thousand Oaks: Sage, 2008. (Quantitative applications in the social sciences, 154).

SATYANARAYANAN, M. The emergence of edge computing. **Computer**, v. 50, n. 1, p. 30-39, Jan. 2017. DOI: https://doi.org/10.1109/MC.2017.9.

SILVERTOWN, J. A new dawn for citizen science. **Trends in Ecology and Evolution**, v. 24, n. 9, p. 467-471, Sept. 2009. DOI: https://doi.org/10.1016/j.tree.2009.03.017.

WILKINSON, M. D.; DUMONTIER, M.; AALBERSBERG, I. J.; APPLETON, G.; AXTON, M.; BAAK, A.; BLOMBERG, N.; BOITEN, J. W.; SANTOS, L. B. da S.; BOURNE, P. E.; BOUWMAN, J.; BROOKES, A. J.; CLARK, T.; CROSAS, M.; DILLO, I.; DUMON, O.; EDMUNDS, S.; EVELO, C. T.; FINKERS, R.; GONZALEZ-BELTRAN, A.; GRAY, A. J. G.; GROTH, P.; GOBLE, C.; GRETHE, J. S.; HERINGA, J.; ´T HOEN, P. A. C.; HOOFT,. R.; KUHN, T.; KOK, R.; KOK, J.; LUSHER, S. J.; MARTONE, M. E.; MONS, A.; PACKER, A. L.; PERSSON, B.; ROCCA-SERRA, P.; ROOS, M.; SCHAIK, R. VAN; SANSONE, S. A.;SCHULTES, E.; SENGSTAG, T.; SLATER, T.; STRAWN, G.; SWERTZ, M. A.; THOMPSON, M.; VAN DER LEI, J.; VAN MULLIGEN, E.; VELTEROP, J.; WAAGMEESTER, A.; WITTENBURG, P.; WOLSTENCROFT, K.; ZHAO, J.; MONS, B. Comment: The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, v. 13, 160018, Mar. 2016. DOI: https://doi.org/10.1038/SDATA.2016.18.

WU, Z.; CHEN, Y.; ZHAO, B.; KANG, X.; DING, Y. Review of weed detection methods based on computer vision. **Sensors**, v. 21, n. 21, 3647, June 2021. DOI: https://doi.org/10.3390/s21113647.