Gerenciando Dados Espaciais Históricos de Pequenas Propriedades Rurais – Caso de Machadinho d'Oeste

Mário Balan, Jaudete Daltio

[1]Empresa Brasileira de Pesquisa Agropecuária (Embrapa)
Soldado Passarinho 303, Fazenda Chapadão – Campinas – SP – Brasil

mario.balan@colaborador.embrapa.br, jaudete.daltio@embrapa.br

Abstract. *Data management in long-term research projects addresses multiple challenges, mostly related to semantic (data interpretation) and structural (collection and storage technologies) heterogeneity. The family farming monitoring project in Machadinho d'Oeste (RO) has these characteristics and aims to monitor 250 variables in 350 small rural properties for 100 years. The purpose of this paper is to present the work in progress which aims to aggregate the project's temporal data with the implementation of a spatiotemporal database. It is expected to assist in the curation of collected data through the aggregation of spatial information and expand the perspective that those involved in the research have on the dynamics of use and occupation of the region.*

Resumo. *A gestão de dados em projetos de pesquisa de longa duração abarca múltiplos desafios, majoritariamente relacionados a heterogeneidade semântica (interpretação dos dados) e estrutural (tecnologias de coleta e armazenamento). O projeto de acompanhamento da agricultura familiar em Machadinho d'Oeste (RO) possui essas características e prevê, por 100 anos, a coleta de 250 variáveis em 350 propriedades rurais. O objetivo deste artigo é apresentar o trabalho em andamento na agregação dos dados históricos do projeto com a implementação de um banco de dados espaço-temporal. Espera-se auxiliar a curadoria dos dados alfanuméricos coletados e expandir a perspectiva que os envolvidos na pesquisa têm sobre a dinâmica de uso e ocupação da região.*

1. Introdução

Projetos de pesquisa de longa duração enfrentam desafios característicos, de múltiplas naturezas, dado a complexidade de suas atividades de gerenciamento (tarefas, pessoas, cronograma) e a constante mudança da equipe envolvida. A gestão de dados nesses projetos apresenta desafios computacionais ainda maiores – tratam-se de dados de uma mesma temática, porém não necessariamente compatíveis, tanto em termos de tecnologia de captura, armazenamento e qualidade, quanto em termos de interpretação [Gonçalves et al. 2018]. O conhecimento acerca do tema da pesquisa também pode evoluir (na verdade, é esperado que isso ocorra), agregando complicadores à produção de resultados baseados na compilação de dados temporais.

O projeto de sustentabilidade e monitoramento da agricultura familiar em Machadinho d'Oeste - RO, desenvolvido pela Embrapa, tem essa natureza [Miranda 2019]. Este projeto teve início na década de 1980 e prevê o acompanhamento, ao longo 100 anos, de

pequenas propriedades rurais no município de Machadinho d'Oeste. Esse acompanhamento é feito em média a cada três anos por meio de entrevistas em campo, baseadas no preenchimento de um questionário com cerca de 250 variáveis sociais, econômicas e agronômicas, que tentam capturar um panorama geral, tão objetivo quanto possível, da realidade dos agricultores.

Juntamente às informações descritivas coletadas (variáveis textuais e numéricas), há um componente espacial associado – cada questionário refere a um lote, delimitado por um polígono. Grandes transformações rurais ocorreram na região desde então e influenciaram diretamente a referência espacial que se tem dos lotes: há lotes abandonados; lotes que foram vendidos e agregados a outros, dando origem a pequenas fazendas; lotes que foram divididos; lotes que incorporaram parte de outros lotes vizinhos.

Até 2008, a coleta de dados era feita em papel por parceiros de instituições locais e as atualizações espaciais eram coletadas de forma rudimentar – era possível inferir alterações espaciais baseando-se na área declarada do lote e no teor do questionário, porém não havia assertividade sobre a alteração ou a nova delimitação [Mangabeira and Grego 2012]. A partir de 2014, implantou-se uma solução com a utilização de dispositivos móveis. Essa migração trouxe inúmeros benefícios, principalmente em termos da acurácia dos dados obtidos, que passaram a contar com a utilização de sensores e com uma autodeclaração espacial do uso da terra.

O objetivo do trabalho apresentado é agregar os dados históricos do projeto Machadinho d'Oeste através da implementação de um banco de dados espaço-temporal que compatibilize os dados das campanhas realizadas. O intuito é conjugar os dados obtidos por meio dos questionários com os dados espaciais e permitir a construção do panorama geral da evolução espacial dos lotes. Por meio deste banco de dados proposto, serão viabilizadas análises que considerem os aspectos espaciais dos dados que, embora não estejam explícitos nos anos anteriores a 2014, são inerentes ao que é declarado nos questionários. Espera-se que a agregação do componente espacial possa auxiliar a curadoria dos dados alfanuméricos coletados e expandir a perspectiva que os envolvidos na pesquisa têm sobre a dinâmica de uso e ocupação da região.

2. Antecedente: Histórico do Projeto

A pesquisa da Embrapa começou a ser concebida em 1982, numa prospecção sobre o processo de assentamento e colonização agrícola de Rondônia. A premissa da pesquisa é validar o projeto de assentamentos implantado na ocasião da definição dos lotes, cujo traçado foi definido de acordo com aspectos topográficos, pedológicos e disponibilidade de água. Uma rede viária foi construída respeitando as curvas de nível, facilitando o acesso às propriedades. O acompanhamento de longo prazo (100 anos) tem o objetivo de caracterizar e monitorar a evolução do uso e ocupação das terras, dos sistemas de produção e da gestão dos recursos naturais praticados. A meta é dimensionar os aspectos relativos à sustentabilidade social, econômica e produtiva de forma a captar elementos de sucesso generalizáveis, obtidos a partir de articulações entre as estratégias locais e as políticas públicas passíveis de aplicação em outras regiões do mesmo bioma.

Embora os indicadores de sustentabilidade agrícola ainda estejam em processo de análise, foi possível obter os números iniciais da campanha de 2018. Apenas 5% dos lotes visitados estavam abandonados, sendo que quase metade dos agricultores entrevis-

tados reside no lote (47%). Cerca de 80% dos entrevistados dedicam mais da metade de seu tempo a atividades agropecuárias no lote. Em termos socioeconômicos, o retorno percebido pelos agricultores é positivo: 74% afirmam que estão melhorando de vida e 70% afirmam não ter a intenção de sair do lote. No cenário agropecuário, a horticultura está presente em 25% dos lotes e a fruticultura em 59%. Entre as culturas anuais e perenes, destacam-se a mandioca e o café. Há uma forte presença de atividades de pecuária, registradas em 25% dos lotes, com destaque para galinhas e bovinos. Mais de 65% dos lotes dedicam parte de sua área para pasto. Os problemas mais recorrentes, apontados por mais de 20% dos agricultores, incluem a baixa fertilidade do solo, documentação de posse e propriedade e ataque de pragas e doenças nas lavouras.

2.1. Evolução dos Mecanismos de Coleta

O acompanhamento tem sido feito de duas formas: remotamente, por meio de imagens de satélite, e *in loco* (em média, a cada três anos), por meio de entrevistas que levantam cerca de 250 variáveis. Até 2008, a coleta de dados em campo era realizada com questionários em papel. Dado o volume de questionários preenchidos (em média 350 preenchimentos), esse procedimento apresentava vários complicadores, desde o tempo gasto na transcrição a potenciais erros de leitura/digitação.

Essa foi a principal motivação para que, a partir da campanha de 2014, fosse adotada uma nova solução que permitisse o preenchimento do questionário em dispositivos móveis [Daltio et al. 2015]. Além dos campos correspondentes ao questionário em papel, foram agregados novos campos oriundos dos sensores dos dispositivos (GPS, câmera e microfone). Imagens de satélite de média resolução espacial com os limites territoriais de cada lote também foram disponibilizadas nos dispositivos para detalhar o uso e ocupação das áreas a partir das declarações dos produtores.

3. Desafios Relacionados ao Projeto

O objetivo de acompanhamento de longo prazo inerente ao projeto implica em desafios relacionados à aquisição, processamento e análise de dados. Passados mais de 30 anos do início do projeto e 10 campanhas de campo, agregar e compatibilizar o volume de dados coletados apresenta inúmeros complicadores. A amostra de lotes entrevistados variou entre campanhas por diversos motivos (estar abandonado ou não ter sido possível encontrar o agricultor que pudesse ser entrevistado). Apesar de tratarem de uma mesma temática, os dados históricos são semântica e estruturalmente heterogêneos.

A heterogeneidade semântica decorre de, a cada campanha, membros serem incluídos e excluídos da equipe do projeto. Novos membros trazem visões complementares e diferentes interpretações sobre o questionário e os dados coletados, associando diferentes níveis de importância a uma mesma pergunta, por exemplo. Essa atuação não homogênea traz reflexos diretos nas decisões tomadas durante a campanha (como proceder quando não é possível prosseguir uma entrevista prevista, por exemplo), no treinamento dos entrevistadores e nas correções do preenchimento do questionário.

Uma soma de fatores contribui para a heterogeneidade estrutural. A primeira delas é relacionada ao próprio questionário. Os campos do questionário também evoluíram ao longo do tempo – algumas perguntas deixaram de ser pertinentes e outras foram inseridas (por exemplo, se agricultor é beneficiado por algum programa social do governo). O teor das respostas também evoluiu, impactando na categorização dos resultados.

Além do questionário em si, a mudança de paradigma de coleta (do papel ao *tablet*) também agregou heterogeneidade tecnológica. Até a campanha de 2008, os dados preenchidos no papel eram transcritos para planilhas, que seguiam as mais variadas estruturas. O processo de transcrição traz erros de interpretação e digitação que precisam ser solucionados pontualmente. A partir de 2014, os questionários já contavam com um campo espacial (a coordenada GPS onde o questionário foi preenchido) e eram agregados em um servidor central, organizados em um banco de dados relacional. Há a necessidade de mapeamento e compatibilização para que essas diferentes bases de dados possam ser analisadas de forma conjunta.

Esses desafios precisam ser considerados na estruturação de um banco de dados espaço-temporal que agregue os dados das campanhas até o momento. Nesse processo será necessário identificar dois pontos essenciais: (i) como identificar univocamente um questionário e sua delimitação espacial (pela criação de uma chave primária, subconjunto de atributos coletados, incorrendo em risco de erros de digitação); e (ii) identificar um subconjunto núcleo de variáveis do questionário, que tenha permanecido semanticamente constante ao longo do tempo e no qual seja coerente realizar análises da evolução histórica. Além disso, a nova estrutura deverá considerar que as potenciais evoluções que ocorrerão no decorrer dos próximos anos e dirimir esforços de grandes reestruturações futuras.

4. Resultados Preliminares

4.1. Aspectos Tecnológicos

Em termos tecnológicos, os dados até a campanha de 2008 estão estruturados em planilhas com diversos *layouts*, de acordo com o questionário seguido em cada campanha. O que se possui de dado espacial é um arquivo vetorial de polígonos em formato shapefile com a delimitação das glebas e lotes amostrados na pesquisa. A partir de 2014, adotou-se como solução o **Open Data Kit** (ODK) [1]. O ODK [Hartung et al. 2010] é um pacote de ferramentas *open-source* composto por: (i) ODK Build (criação dos formulários); (ii) ODK Collect (coleta de dados); e (iii) ODK Aggregate (gerenciamento centralizado dos dados coletados). O SGBD *PostgreSQL* com sua extensão espacial *PostGIS* foi utilizado para armazenar os dados obtidos pelo ODK.

4.2. Análise do Questionário e sua Abrangência Espacial

As primeiras análises foram realizadas nos dados das últimas campanhas (2014 e 2018) e têm o objetivo de identificar univocamente um questionário e seu escopo espacial (a qual polígono esse questionário se refere). O intuito é elencar alterações significativas que possam direcionar refinamentos específicos para os demais anos. As análises partem da tentativa de identificar um conjunto de dados utilizando como chave a tupla gleba + lote. Nas campanhas de 2014 e 2018, os entrevistadores foram orientados a: (i) realizar a entrevista dentro do lote e (ii) capturar a coordenada geográfica pelo GPS do *tablet* junto ao questionário no momento da entrevista. Partindo dessa premissa, a primeira análise realizada foi verificar se o ponto coletado estava de fato dentro dos limites esperados para a propriedade. O parâmetro de junção não espacial considerado para identificar inconsistências foi a tupla gleba + lote preenchida textualmente no questionário.

[1] opendatakit.org

Para esta verificação, uma camada de pontos contendo as coordenadas obtidas na pesquisa em campo foi sobreposta a uma camada de polígonos contendo os loteamentos originais. Em seguida, foi determinada uma área de influência de 150 metros para os pontos da primeira camada, cuja dissociação com os polígonos da segunda camada resultou no conjunto de pontos selecionados para uma análise pormenorizada. A área de influência foi determinada em 150 metros devido a (i) precisão do GPS dos *tablets*, que variou entre 4 e 16 metros, (ii) a ausência da representação de vias públicas na camada de polígonos, o que ocasiona uma pequena distorção, (iii) o tamanho médio dos lotes (50 ha), pouco inferior ao módulo fiscal do município (60 ha), e (iv) o traçado de lotes mais recorrente, que possui cerca de 300m de frente. O objetivo foi encontrar as coordenadas de locais de preenchimento de questionário que estivessem significativamente distantes do lote a que se referiam.

Nos dados referentes a 2014 notamos 12 ocorrências (de 384 questionários); em 2018 foram 11 ocorrências (de 414). A Figura 4.2 ilustra dois desses casos de potenciais inconsistências: em 2014, o questionário do lote 445 foi preenchido em ponto na proximidade da divisa entre os lotes 422, 396 e 405. Em 2018, o questionário do lote 233 foi preenchido na proximidade da divisa do lote 180 e 234. Essas inconsistências podem ter mais de uma interpretação e precisam ser analisadas. O que buscamos identificar são casos em que o questionário foi preenchido dentro de outro lote, o que de fato nos leva a alterar a sua referência espacial (a gleba + lote que descreve). A partir dessa verificação é possível inferir a dinâmica de agregações e desmembramentos dos lotes pesquisados.

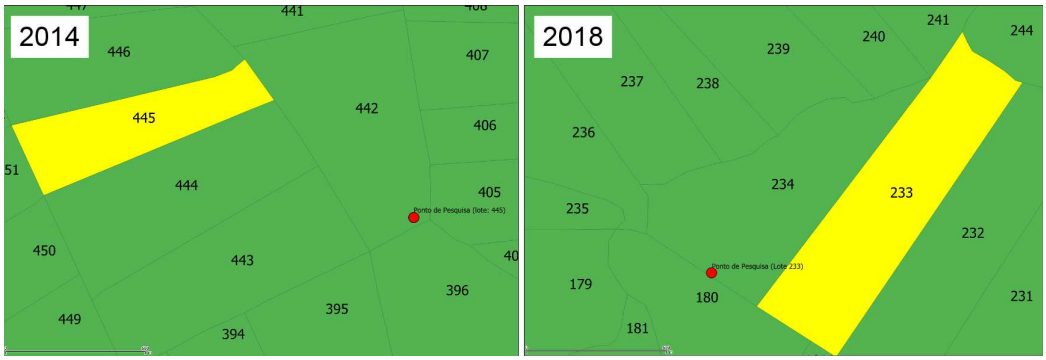

Figura 1. Abrangência Espacial: ponto de pesquisa *versus* polígono do pote

4.3. Análise da Evolução do Uso e Ocupação dos Lotes

Nas pesquisas de 2014 e 2018 foi incorporada ao questionário uma representação cartográfica do lote (um mapa) sobreposta a uma imagem de satélite recente. Baseado nessa imagem, o entrevistador identificou e delimitou, de acordo com a declaração do agricultor, os principais elementos de uso e ocupação propriedade pesquisada: áreas construídas, pasto, culturas e vegetação nativa, por exemplo.

Essas anotações estão em processo de espacialização pela equipe do projeto e irão permitir uma avaliação da evolução espaço-temporal referente ao uso e ocupação dos lotes, além de subsidiar a validação da área espacial a qual o questionário se refere, também o faz para as variáveis presentes no questionário, como a área (em termos numéricos) dedicada ao cultivo de culturas anuais, perenes, área dedicada a pasto ou área de mata nativa

preservada dentro do lote. A Figura 4.3 ilustra um caso de evolução, onde é possível verificar o deslocamento da área dedicada ao cultivo de mandioca, que em 2018 substituiu a área dedicada ao cultivo café, e o avanço da área de pastagem.

Figura 2. Evolução do uso e ocupação autodeclarado entre 2014 e 2018

5. Contribuições e Trabalhos Futuros

O projeto de sustentabilidade e monitoramento da agricultura familiar em Machadinho d'Oeste apresenta inúmeros desafios computacionais no que tange a gestão dos dados. Esses desafios têm sido enfrentados gradativamente pelas equipes que, rotativamente, vêm contribuindo com o projeto. A evolução do mecanismo de coleta de papel para *tablet* trouxe ganhos expressivos em termos de eficiência nos trabalhos e o risco de perda de informações foi praticamente eliminado. O uso de sensores contribuiu ainda mais para que os dados obtidos nas entrevistas de 2014 e 2018 tivessem qualidade e confiabilidade superior aos das campanhas anteriores e expandiu consideravelmente as possibilidades de uso das informações espaciais como mecanismos de validação e compreensão dos dados coletados via questionário. Os trabalhos futuros preveem a continuidade das atividades de integração dos dados: a validação da delimitação dos lotes; a espacialização das anotações espaciais de uso e ocupação; e a identificação de um subconjunto de variáveis semanticamente coeso em todas as campanhas (1986 a 2018).

Referências

Daltio, J., Martinho, P. R. R., Magalhães, L. A., and Carvalho, C. A. (2015). Utilização de Dispositivos Móveis para Coleta de Dados em Campo - Experiência Machadinho d'Oeste. In *X SBIAGRO*, pages 1078–1091, Ponta Grossa, RS.

Gonçalves, G. C., Augusto, L., Escada, M. I. S., and Amaral, S. (2018). Spatial database to store years of earth observation information obtained fromfield expeditions in the amazon. In *Proceedings XIX GEOINFO*, pages 134–139, Campina Grande, PB.

Hartung, C., Lerer, A., Anokwa, Y., Tseng, C., Brunette, W., and Borriello, G. (2010). Open Data Kit: Tools to Build Information Services for Developing Regions. In *Proc. 4th ACM/IEEE ICDT*, pages 18:1–18:12, New York, USA.

Mangabeira, J. A. C. and Grego, C. R. (2012). Café com Leite: o perfil dos produtores rurais de Machadinho d'Oeste, RO, em 2008. In *Documentos Embrapa, 96*.

Miranda, E. E. (2019). Sustentabilidade Agrícola na Amazônia - Machadinho d'Oeste. Disponível em www.machadinho.cnpm.embrapa.br. Embrapa.