Preparação de dados para modelagem de processos logísticos agrícolas

Palavras-Chave: produtos agrícolas, banco de dados, custo de transporte.

Autores/as:
Thiago Tribiolli Bazzi (aluno) [Faculdade de Engenharia Agrícola]
Prof.ª Dr.ª Andréa Leda Ramos de Oliveira (orientadora) [Faculdade de Engenharia Agrícola]

INTRODUÇÃO:

O Brasil se encontra entre os principais países exportadores de produtos agropecuários, ocupando a quarta posição, com aproximadamente US$ 96,9 bilhões em exportações. Dentre os principais produtos produzidos e exportados temos a soja, carne bovina, milho e café (CNA, 2020).

Um desafio encontrado ainda é com relação a processos logísticos, que executados de forma ineficiente geram perdas significativas dos ganhos de produção, influenciando na competitividade do mercado nacional e internacional do agronegócio brasileiro (OLIVEIRA, 2014; EMBRAPA, 2018). O sistema rodoviário é o principal modal de transporte, integrando as principais regiões produtoras ao mercado doméstico e portos exportadores. Neste sentido, o conhecimento do comportamento dos fretes rodoviários agrícolas é uma importante ferramenta de apoio para tomada de decisões. (MOREIRA et al., 2017)

A modelagem de sistemas logísticos, como modelos matemáticos e simulações computacionais, é desenvolvida pensando na otimização do escoamento dos produtos, e busca refletir no maior rendimento em relação a custos e lucros. A modelagem utiliza-se de uma série de ações organizadamente planejadas para realização de um modelo operacional. Entre essas ações temos a coleta e preparação de dados. Segundo Junqueiro e Morabito (2006), a utilização da modelagem auxilia nas tomadas de decisão referentes ao planejamento da cadeia de produção, processos de estocagem, meios de transporte e outros, que visam minimizar gastos e custos logísticos, potencializando o rendimento econômico da produção.

O processo de mineração de dados é um método que pode trazer uma estimada recompensa a todas as áreas, como a logística agroindustrial. Com este método podemos identificar padrões de novas tendências e melhorias nos processos que envolvem o campo da atuação. O *Knowledge Discovery in Databases*, extração de conhecimento de banco de dados, busca constantemente novos métodos a serem implementados para atingir cada vez mais a eficiência dos dados, etapa que antecede processos como a modelagem. Sua constituição é de três passos básicos, que são o pré-processamento (preparação), a *Data Mining* (Mineração de Dados) e o pós-processamento ou interpretação dos resultados (DANTAS, 2008).

A preparação de dados em projetos e pesquisas de Descoberta de Conhecimento em Bancos de Dados (DCBD) consomem cerca de 60% do tempo de execução. De acordo com Pyle (1999), a fase de preparação de dados é a mais longa das fases da mineração de dados, pois envolve etapas que vão desde a identificação do problema até a modelagem e formulação da apresentação dos resultados. Compreendendo a verificação de erros, identificação de outliers, programação fazendo uso de métodos lógicos e tecnologias, verificação e validação por análises estatísticas e simulações.

A disponibilidade de uma grande quantidade de dados e de métodos computacionais abrem diversas possibilidades em modelagem. Assim, esse projeto visa contribuir com as pesquisas já existentes, através da coleta, preparação dos dados e análises exploratórias para posterior aplicação

de técnicas e métodos de modelagem. Esperando promover a iniciação na ciência por meio de procedimentos universais que se aplicam em diversas áreas do conhecimento.

OBJETIVO:

O projeto tem como objetivo geral preparar e explorar um banco de dados de fretes logísticos de produtos agrícolas selecionados para aplicação de técnicas de mineração de dados. Construir superfícies de custos de transporte para todo o território brasileiro.

MATERIAIS E METODOLOGIA:

A identificação dos dados de custos de transportes foi obtida na base do Laboratório de Logística e Comercialização Agroindustrial (LOGICOM), que reúne os Anuários do Sistema de Informações de Frete (SIFRECA) do grupo ESALQ-LOG, dados da Confederação Nacional dos Transportes (CNT) e do Instituto Mato-grossense de Economia Aplicada (IMEA). Como forma de execução, para a preparação dos dados, foi utilizado a ferramenta Microsoft Excel, *software* de criação e edição de planilhas eletrônicas, e para a construção da superfície de custos foi utilizada a ferramenta ArcGis, *software* de informação geográfica. Para a execução deste projeto, os métodos foram divididos em cinco etapas principais: identificação dos dados, preparação dos dados, outras localidades, análise estatística exploratória e construção da superfície de custos de R$/t.km.

RESULTADOS E DISCUSSÃO:

Com a identificação e posteriormente com o tratamento e preparação dos dados, foi possível determinar, por ano, a quantidade total dos dados e deste total, quais eram as localidades de municípios e não municípios, conforme mostrado na Tabela 1.

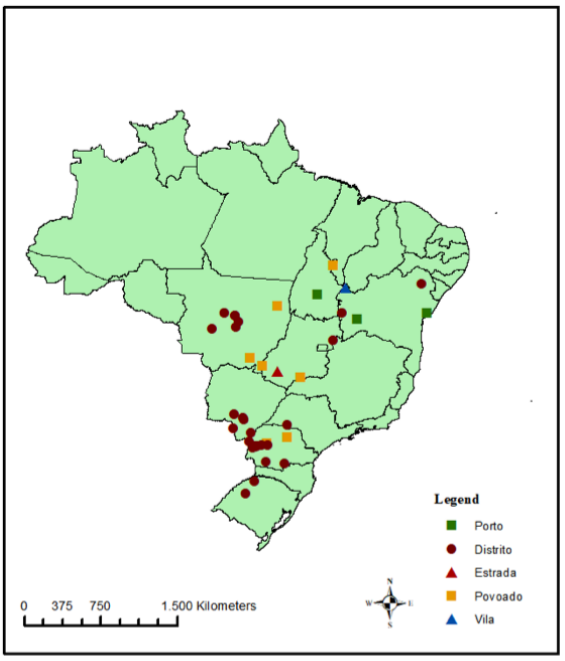
Tabela 1 - Quantidade de dados do banco de dados trabalhado por ano estudado.

Ano	Total	Municípios	Não Municípios	Excluídos (*Boxplot*)
2012	7222	6861	361	261
2013	6685	6358	327	216
2014	5050	4884	166	179
2015	3108	3032	76	79
2016	2391	2306	85	31
2017	2305	2177	128	26
2018	2655	2553	102	84
2019	2483	2425	58	66
2020	2241	2197	44	58

Fonte: SIFRECA: Sistema de Informação de Frete. ESALQ-LOG: Grupo de Pesquisa e Extensão em Logística Agroindustrial. IMEA: Instituto Mato-grossense de Economia Aplicada. CNT: Confederação Nacional dos Transportes.

Com a preparação e organização dos dados, foi possível, também, realizar a identificação das localidades que não eram municípios, sendo portos, distritos, estradas, povoados ou vilas, como mostrado na Figura 1, permitindo a visualização em território nacional de suas respectivas unidades federativas.

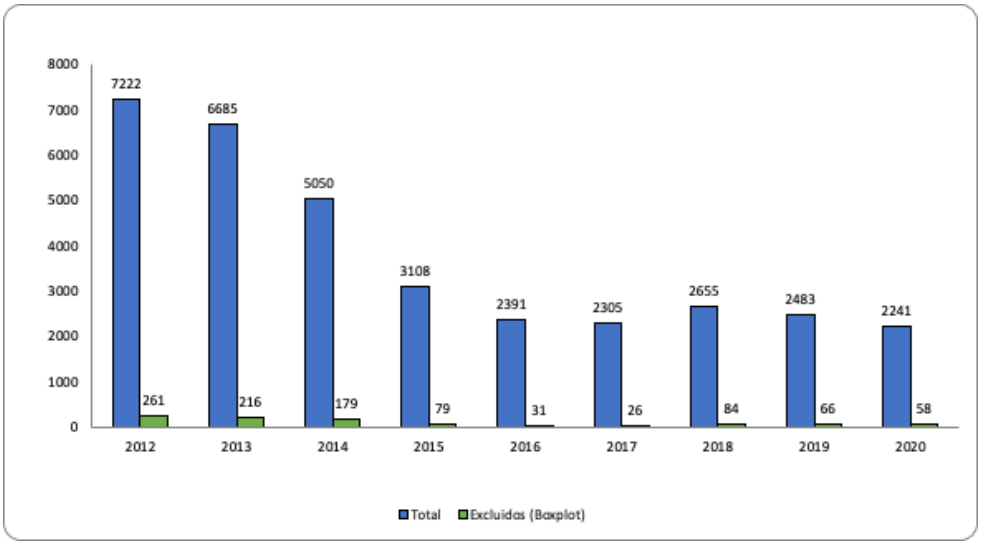
Figura 1 - Localização de outras localidades em território nacional, para os anos de 2012, 2016 e 2019.



Fonte: Elaborado pelo autor.

Com relação aos dados das localidades dos municípios, que passaram pelo mesmo tratamento inicial aos dados das localidades de não municípios, ocorreu há análise estatística exploratória, pelo gráfico *boxplot* do método dos *outliers* distantes de Tukey, referentes aos valores de R$/t.km para construção da superfície de custos. Os resultados obtidos podem ser visualizados na Figura 2, observando que o número de dados excluídos tem relação direta com a quantidade total de dados, conforme mostrado também na Tabela 1.

Figura 2 - Relação da quantidade total de dados por dados excluídos (*boxplot*), no período de 2012 até 2020.

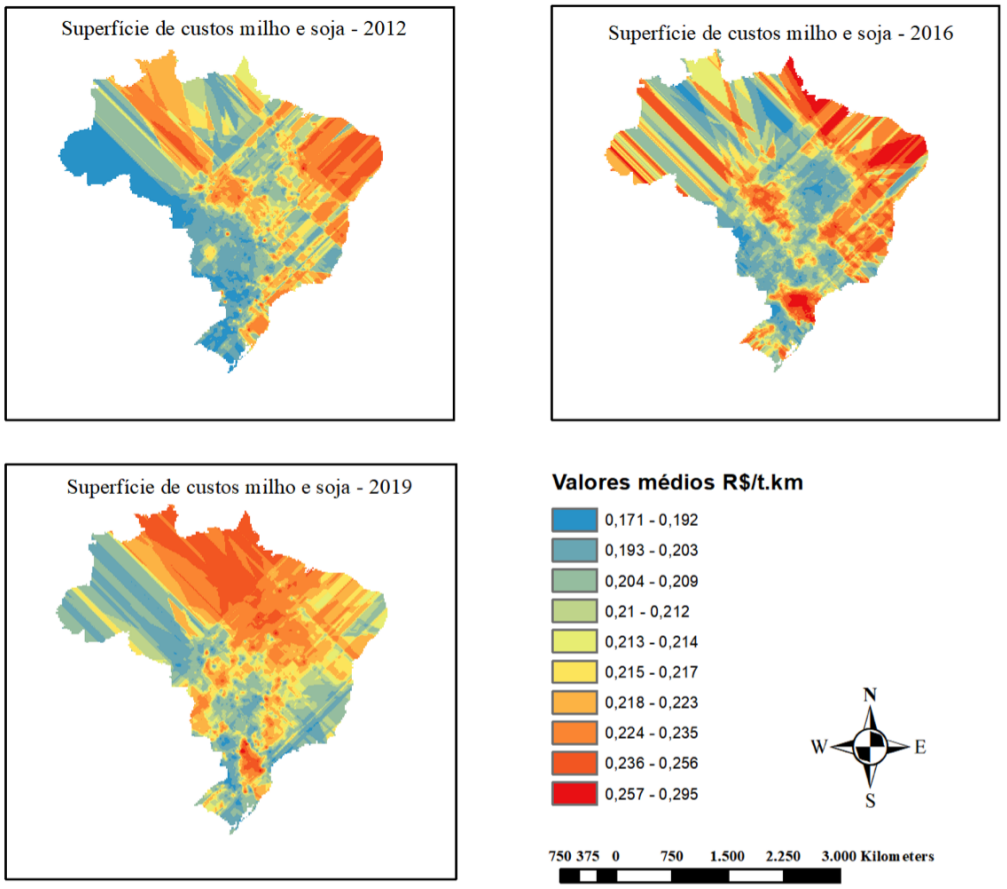


Fonte: Elaborado pelo autor.

Assim como a identificação e o tratamento dos dados de outras localidades, os resultados da superfície de custos dos valores médios de R$/t.km corroboram para a delimitação de bacias

logísticas para a agropecuária brasileira, conforme visto em Castro et al. (2019) - Macrologística: Caracterização das bacias logísticas da agropecuária brasileira. Os resultados podem ser observados na Figura 3, onde os valores médios de R$/t.km, após passagem pela exclusão de dados *outliers* do método *outlier* distantes utilizado, se encontram no intervalo de 0,171 até 0,295. Nos valores mais baixos, representados pela coloração azul, encontram-se as principais regiões produtoras de grãos do país, regiões Centro-Oeste, Sudeste e Sul, onde se encontra, também, a maior quantidade de origem dos dados.

Figura 3 - Superfície de custos rodoviários R$/t.km para 2012, 2016 e 2019.



Fonte: Elaborado pelo autor.

CONCLUSÕES:

Os resultados corroboram para aplicações de estudos e pesquisas em logística agroindustrial, sendo aplicável em outras áreas. Com a identificação das imprecisões foi possível determinar métodos, utilizando de ferramentas, que auxiliam no desenvolvimento de outros projetos de pesquisa. É fundamental a realização constante de progresso e estímulos na geração de novos métodos dentro da ciência de preparação de dados. Sendo esta uma prática considerada crítica e demorada, podendo consumir até 80% do tempo de analistas e pesquisadores convertendo dados brutos em informações de alta qualidade e prontas para análise.

Todos os processos desenvolvidos na metodologia neste projeto de pesquisa fazem uso em grandes bancos de dados. Foram utilizados 34.140 dados no total, passando por todas as partes necessárias de análise e correções. Dentro do tratamento de dados a utilização da ciência e matemática estatística é de caráter crucial para o propósito de qualquer pesquisa e projeto. A análise

exploratória de dados utiliza técnicas estatísticas univariadas para identificar padrões ou tendências que podem estar ocultos em dados agrupados. Essa análise preliminar favorece a avaliação da qualidade dos dados coletados.

Durante todo o desenvolvimento e com a confirmação dos resultados obtidos, as informações, para qualquer pesquisa, projeto ou estudo, necessitam estar bem preparadas para realização do objetivo final. Assim, o tratamento de dados é uma atividade essencial de autoatendimento que converte dados diferentes, brutos e desorganizados em uma visualização limpa e consistente. Este processo inclui procura, limpeza, transformação, organização e coleta de dados.

BIBLIOGRAFIA

CASTRO, G. S. A.; CARVALHO, C. A; DALTIO, J.; MIRANDA, E. E.; MAGALHÃES, L. A. Macrologística: caracterização das bacias logísticas da agropecuária brasileira. **Anais do XIX Simpósio Brasileiro de Sensoriamento Remoto**, INPE – Santos-SP, Brasil. 2019.

DANTAS ERG, et al. O uso da descoberta de conhecimento em base de dados para apoiar a tomada de decisõe**s**. **V Simpósio Excel em Gestão e Tecnol**. 2008; 1–10.

EMBRAPA; Empresa Brasileira de Pesquisa Agropecuária Embrapa Territorial Ministério da Agricultura, Pecuária e Abastecimento. **Macrologística da Agropecuária Brasileira: Delimitação das Bacias Logísticas**. Estudos Logísticos, [2018]. Disponível em: https://www.embrapa.br/macrologistica/estudos-logisticos. Acesso em: 14 de abr. 2020.

Instituto Brasileiro de Geografia e Estatística - IBGE. **Códigos dos Municípios.** Disponível em: < https://www.ibge.gov.br/explica/codigos-dos-municipios.php>. Acesso em: 10 de out. 2020.

JUNQUEIRA, R. A. R.; MORABITO, R. Um modelo de otimização linear para o planejamento agregado da produção e logística de sementes de milho. **Produção**, v. 16, n. 3, p. 510-525, 2006.

MOREIRA, C. E. S.; OLIVEIRA, A. L. R.; OLIVEIRA, S. R. M.; YAMAKAMI, A. Identification of freight patterns via association rules: the case of agricultural Grains. **Bulgarian Journal of Agricultural Science**, v. 23, n. 6, p. 887-893, 2017.

PANORAMA do agro, **Confederação da Agricultura e Pecuária do Brasil**, 2020. Disponível em: <https://www.cnabrasil.org.br/cna/panorama-do-agro>. Acesso em: 16 fev. 2021.

PYLE, D. **Data Preparation for Data Mining**. São Francisco: Morgan Kaufmann, 1999. 466p.

Sistema de Informações de Fretes - SIFRECA. **Fretes Rodoviários**. Disponível em: <http://sifreca.esalq.usp.br/mercado-de-fretes/soja/>. Acessado em: 10 out. 2020.

Tukey JW. *Exploratory Data Analysis*. **Addison-Wesley**, preliminary edition, 1970.