Avaliação de Dimensões de Qualidade de Dados para o Agronegócio

Clovis S. Junior[1], **Carina F. Dorneles**[2]

[1]Instituto de Ciencias Exatas e Naturais - ICEN
Universidade Federal de Rondonólis - UFR/Rondonópolis-MT

[2]Departamento de Informatica e Estatística - INE
Universidade Federal de Santa Catarina - UFSC/Florianópolis-SC

clovis@ufr.edu.br, carina.dorneles@ufsc.br

Abstract. *Good quality data improves information accuracy resulting in assertive decisions. To assess the various aspects involved, an approach that makes use of dimensions for verification is important. In this article, a study about data quality dimensions in the agribusiness domain is presented. The validation of dimensions is carried on two real databases, the first focusing on the environment and the second on family farming. The results show how each database behaves with defined dimensions.*

Resumo. *Dados de boa qualidade melhoram a precisão de informações resultando em decisões mais assertivas. Para avaliar os diversos aspectos envolvidos, é necessária uma abordagem que faça uso de dimensões para a verificação. Neste artigo, é apresentado um estudo acerca de dimensões de qualidade de dados no domínio do agronegócio. A validação das dimensões é feita sobre duas bases de dados reais, a primeira com foco no meio ambiente e a segunda em agricultura familiar. Os resultados mostram como cada base de dados se comporta em relação às dimensões definidas.*

1. Introdução

O conceito de qualidade dos dados não é único e pode ser muito subjetivo. Uma definição mais ampla e consensual remete à adequação ao uso, e pode se basear no contexto pois os dados podem ser considerados adequados para um cenário e não ser apropriados para outro. Outra definição considera multidimensionalidade, que pode ser considerado como ”atributos que permitem representar uma característica particular” [Sidi et al. 2012]. As dimensões da qualidade podem ser referidas tanto aos valores dos dados quanto ao seu esquema. Embora a qualidade dos esquemas seja reconhecida como uma área de pesquisa relevante, a maioria das definições de dimensões e métricas de qualidade de dados são referentes a valores de dados [Batini et al. 2009].

Diversos trabalhos na literatura analisam questões de qualidade de dados, focados em domínios específicos [Cichy and Rass 2019, Malaverri and Medeiros 2012]. Por exemplo, em [Cichy and Rass 2019], são analisados pontos quanto à definição, avaliação e melhoria dos dados, porém com foco em metodologias para ambientes de negócios. No entanto, além da abordagem não estar relacionada com o agronegócio, a análise abrange poucas dimensões. [Malaverri and Medeiros 2012] apresentam uma revisão a respeito de

qualidade de dados direcionada à agricultura e ciências geoespaciais. No trabalho, os autores fazem algumas análises quanto a formas de validar dimensões de qualidade de dados classificando-as em análises manuais e automatizadas. Os aspectos abordados são especificamente agrícolas, diferentemente das investigações apresentadas neste artigo, cuja abordagem foca no agronegócio.

A abordagem apresentada neste artigo foi desenvolvida a partir da definição de dimensões de forma mais abrangente com dimensões identificadas na literatura, sem domínio específico. Em seguida, foi realizado um refinamento por profissionais especialistas na área. O artigo também apresenta uma abordagem prática e empírica [Sadiq and et al. 2018] feita através de simulações para cada dimensão indicada pelos especialistas. A proposta é identificar as dimensões de qualidade de dados relevantes para o agronegócio e utilizá-las, em trabalhos futuros, como indicadores de quais atributos demandam enriquecimento de dados.

2. Trabalhos relacionados

Nesta seção, são apresentados trabalhos para identificação inicial das dimensões de qualidade de dados, sem distinção de área. Estas dimensões foram especializadas para o domínio de agronegócio por meio de investigação junto a especialistas da área. A gestão da qualidade de dados tem se tornado importante tanto do ponto de vista acadêmico quanto sob perspectivas profissionais devido à crescente consciência dos impactos causados por dados pobres. A pesquisa proposta por [Silvola et al. 2016] faz uma abordagem integrando dimensões diferentes incluindo avaliação cognitiva. O trabalho apresentado em [Sidi et al. 2012] aborda qualidade de dados como suporte a vários serviços organizacionais.

A falta de qualidade de dados nas organizações pode resultar em diversos problemas em sistema de informação cooperativo. A pesquisa proposta por [Cichy and Rass 2019] fornece uma visão geral de estruturas completas de qualidade de dados que são amplamente aplicáveis, resumindo e comparando seus componentes principais, incluindo a definição de qualidade de dados, avaliação e processos de melhoria. O artigo apresentado por [Malaverri and Medeiros 2012] fornece uma base conceitual para o desenvolvimento de aplicações na agricultura. Dados em aplicações agrícolas podem ser temáticos, textuais ou geoespaciais. A pesquisa não analisa apenas questões relativas à qualidade de dados geoespaciais, considera a qualidade em todos os tipos de dados e fornece diretrizes para aplicações agrícolas. Qualidade de dados em arquiteturas de *big data* inclui propriedades tais como volume referente à grande quantidade de dados (normalmente em TB ou magnitudes acima), velocidade que os dados são formados, e variedade, indicando que *big data* tem todos os tipos de dados estruturados e não estruturados [Cai and Zhu 2015].

3. Dimensões de Qualidade de Dados para o Agronegócio

A identificação inicial das dimensões de qualidade de dados foi feita com revisão da literatura sem distinção de área. Foram identificadas 54 dimensões, posteriormente foram especializadas através de um questionário simples, junto a pessoal técnico relacionado com tecnologia da informação para o agronegócio como programadores, gerentes de projetos e gerentes agrícolas. A etapa seguinte foi realizada pelos colaboradores através da

avaliação empírica das dimensões e sua importância dentro do agronegócio. Posteriormente, foi feito a identificação das dimensões mais relevante para o contexto. A terceira e última etapa foi a avaliação das dimensões de qualidade, realizada através de simulações e disponibilizando os resultados através de gráficos.

3.1. Identificação das dimensões

Na interface do questionário, os participantes foram instruídos a indicar sua percepção de relevância para cada dimensão como requisito de qualidade de dados para o agronegócio. A relevância foi verificada utilizando uma escala entre 1% e 100%, para as 54 dimensões. Ao final, utilizou-se um ponto de corte de relevância$>= 90\%$. O objetivo foi identificar as dimensões adequadas para o agronegócio, descartando dimensões genéricas de acordo com os colaboradores. De acordo com os especialistas, foram identificadas 10 dimensões adequadas aos parâmetros definidos entre as 54 investigadas.

- Atualização: Ocorre se os dados estiverem corretos apesar de discrepâncias de tempo.
- Auditabilidade: É observado a possibilidade dos auditoria da integridade dos dados.
- Disponibilidade: Grau no qual a informação é fisicamente acessível.
- Credibilidade: Ponto em que os dados são considerados verdadeiros e confiáveis.
- Consistência: Como os dados são apresentados e compatíveis com versões anteriores.
- Fund. de integridade de dados: Grau de existência da obrigatoriedade de atributos.
- Especificação de dados: Nível de existência, integridade, modelos de dados e regras de negócios.
- Eficiência: Capacidade dos dados em atender rapidamente às necessidades de determinada tarefa.
- Integridade: Clareza e critérios pré-estabelecidos como consistência, integridade estrutural e de conteúdo.
- Validade: Conformidade com a sintaxe de sua definição ou seu propósito.

4. Avaliação e Resultados

A validação das dimensões foi realizada sobre duas bases de dados reais, a primeira com foco no meio ambiente e a segunda referente a dados de propriedades rurais destinadas à agricultura familiar. **Bd-Agricultura (ag)**. base de dados utilizada no projeto "Sistema para Coleta de Dados Socioeconômicos em Comunidade Rurais", o objetivo é coletar dados e fornecer informações para tomada de decisões em nível municipais. A base de dados possui 74 tabelas e um volume de 13,19 MB de dados. **Bd-Ambiental (am)**. base de dados do projeto "Sistema para Gerenciamento de Planos de Recuperação em Áreas Degradadas, como estratégia de consolidação do Novo Código Florestal Brasileiro", o objetivo é o gerenciamento de Planos de Recuperação de Áreas Degradadas (PRAD). A base de dados possui 62 tabelas e um volume de 8,25 MB de dados.

4.1. Métricas de avaliação

As métricas usadas para avaliação foram definidas considerando cada uma das dimensões definidas. Os parâmetros usados foram extraídos de cada uma das bases de dados, considerando as especificações em seus metadados.

Atualização: verifica a média temporal de atualizações dos atributos para um determinado esquema, $SomaAtualizaColunas_{emdias}$ = soma da qtde de colunas atualizadas, $totalTabelas$ = qtde total de tabelas do esquema:

$$Atualizacao = SomaAtualizaColunas_{emdias}/totalTabelas \quad (1)$$

Auditabilidade: verifica o percentual de tabelas auditáveis frente ao total de tabelas no esquema, $totalTabAud_{meta}$ = total de tabelas auditáveis(metadados), $totalTabelas$ = total de tabelas do esquema:

$$Auditabilidade = (totalTabAud_{meta}) * 100/totalTab \quad (2)$$

Disponibilidade: avalia elementos não obrigatórios, apresentando o percentual de colunas obrigatórias em relação à qtde no esquema, $totalColObrig$ = qtde total de colunas obrigatórias, $totalCol$ = total de colunas do esquema:

$$Disponibilidade = (totalColObrig * 100)/totalCol \quad (3)$$

Credibilidade: verifica se valores textuais (nomes) pertencem a um conjunto de 201.574 nomes válidos, $totalDados_{dicNomes}$ = qtde de dados (conjunto), $DadosBD$ = total de dados armazenados, $totalLinhas$ = total de dados no BD:

$$Credibilidade = (((totalDados_{dicNomes}) * 100)/DadosBD)/totalLinhas \quad (4)$$

Consistência: verificação é realizada de forma cronológica para permitir uma verificação comparativa. A simulação deve ser repetida em momentos diferentes, $tamColuna_{IFtamUsado=tamDef}$ = qtde de armazenamento médio usado, $totalLinhas$ = total de dados no BD:

$$Consistencia = ((tamColuna_{IFtamUsado=tamDef}/totalLinhas) * 100)/totalLinhas \quad (5)$$

Fundamentos de integridade de dados: verifica o percentual de atributos não obrigatório em relação ao todo. Poucos atributos obrigatórios favorecem baixo preenchimento de dados e baixa integridade, $totalLinhas_{colNoObrig}$ = total de colunas não obrigatórias, $totalLinhas$ = qtde total de linhas de dados no BD:

$$Fundamentos = ((totalLinhas_{colNoObrig}/totalLinha) * 100)/totalLinhas \quad (6)$$

Especificação de dados: verifica a quantidade de restrições em relação à quantidade de tabelas, poucas restrições indicam preenchimentos incorretos de dados, $totalRestricoes$ = qtde de restrições nas tabelas do esquema, $totalTabelas$ = total de tabelas do esquema:

$$Especificacao = ((totalRestricoes) * 100)/totalTabelas \quad (7)$$

Eficiência: verifica a legibilidade para definição de atributos. Foi usado um conjunto de 261.798 de palavras verificando a legibilidade dos atributos nos metadados do domínio, $totalColunas_{dicNomes}$ = total de nomes no conjunto de dados, $totalLinhas$ = total de linhas de dados no esquema:

$$Eficiencia = (((totalColunas_{dicNomes})/totalLinhas) * 100))/totalLinhas \quad (8)$$

Integridade: o cálculo é feito a partir da média de dados sem nomes abreviados em relação a quantidade de linhas de dados em todas as tabelas do esquema, $totalColuna_{dadosNoAbrev}$ = total de nomes não abreviados, $totalLinhas$ = total de dados no esquema:

$$Integridade = ((totalColuna_{dadosNoAbrev}/totalLinhas) * 100)totalLinhas \quad (9)$$

Validade: foram utilizados 2 atributos para simulação: Nome>10 elementos e CPF=11 elementos, $totalLinhas_{nomeCpf}$ = total de dados (CPFs e Nomes), $totalLinhas$ = total de dados:

$$Validade = (((totalLinhas_{nomeCpf})/totalLinhas) * 100)/totalLinhas \quad (10)$$

4.2. Resultados

Nesta seção, são apresentados os resultados da avaliação apresentados por meio de gráficos comparativos entre as duas bases de dados, ag e am.

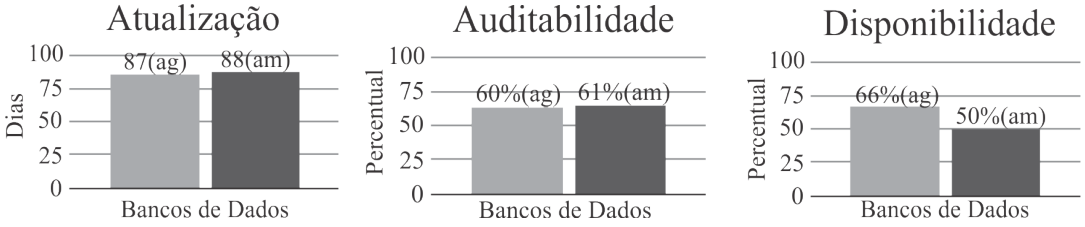

Figura 1. (a) Atualização; (b) Auditabilidade; (c) Disponibilidade

A Figura 1 apresenta o resultado das dimensões atualização, auditabilidade e disponibilidade. Atualização apresenta dados semelhantes indicando em média 87(ag) e 88(am) dias sem atualizações[1], poucas atualizações indicam empobrecimento de dados. Auditabilidade os resultados foram próximos, 60%(ag) e 61%(am). Tabelas auditáveis dependem de estratégia, mas, grandes quantidades de tabelas auditáveis permitam rastrear inconsistências. Disponibilidade resultou em 66%(ag) e 50%(am) para colunas obrigatórias, destacando que, atributos obrigatórios (NOT NULL) impõem a existência do dado.

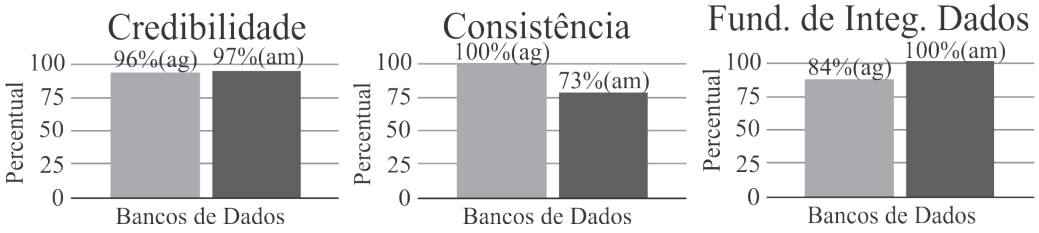

Figura 2. (a) Credibilidade; (b) Consistência; (c) Fund. de integridade de dados

A Figura 2 apresenta o resultado para as dimensões credibilidade, consistência e fundamentos de integridade de dados. A credibilidade apresenta percentualmente dados para armazenamento de nomes próprios quanto a legibilidade, outros critérios podem ser utilizados, os resultados para essa dimensão foram próximos, 96%(ag) e 97%(am). Consistência apresentou resultados percentuais de 100%(ag) e 73%(am)[2]. Fundamentos de integridade de dados resultaram em 34%(ag) e 50%(am), com atributos não obrigatórios indicando possíveis problemas.

[1]As 2 bases não estavam em atividade na fase de testes; usualmente, são atualizadas diariamente.

[2]Consistência verifica atributos textuais, BD-Ambiental é majoritariamente numérico resultando em percentuais baixos

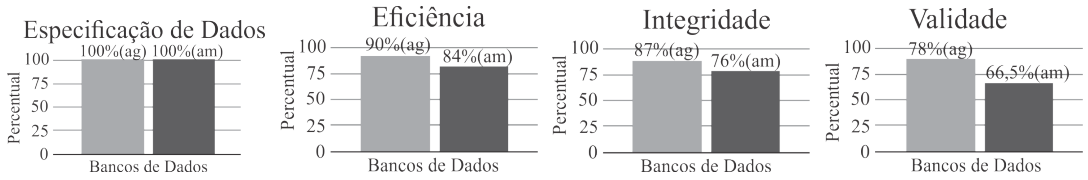

Figura 3. (a) Especificação de dados; (b) Eficiência; (c) Integridade; (d) Validade

A Figura 3 apresenta o resultado para as dimensões especificação de dados, eficiência, integridade e validade. Especificação de dados apresentou percentuais em 100% para as duas bases de dados indicando ao menos uma restrição (*constraint*). Eficiência analisa atributos com nomenclaturas legíveis. Atributo como CP001 não fornece referência em comparação a atributos como ENDEREÇO etc, os percentuais obtidos foram 10%(ag) e 4%(am) indicando pouca eficiência. Integridade é apresentada percentualmente de acordo com a quantidade de nomes abreviados (nomes ou descrições), os resultados 87%(ag) e 76%(am) indicam poucas abreviações, favorecendo a integridade. Os resultados foram 22%(ag) e 10.5%(am) indicando pouco controle na qualidade dos dados inseridos.

5. Conclusões e Trabalhos futuros

Mensurar a qualidade de dados, implica em entender as particularidades de cada domínio, com isso, é necessário identificar quais dimensões são relevantes. Esse artigo identificou dimensões de qualidade de dados para o agronegócio com auxílio técnico de profissionais de diferentes vertentes. Foram apresentas dimensões para o agronegócio com simulações práticas, realizadas com bases de dados reais, o propósito foi exemplificar o uso das dimensões. Como trabalhos futuros, as dimensões investigadas no artigo servem como referência para criação de abordagens para enriquecimento de dados para o agronegócio com extensões para visualização de dados.

Referências

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. 41(3).

Cai, L. and Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14:2.

Cichy, C. and Rass, S. (2019). An overview of data quality frameworks. *IEEE Access*.

Malaverri, J. and Medeiros, C. (2012). Data quality in agriculture applications. *Proceedings of the Brazilian Symposium on GeoInformatics*, pages 128–139.

Sadiq, S. and et al. (2018). Data quality: The role of empiricism. *SIGMOD Rec.*, 46(4):35–43.

Sidi, F., Shariat Panahy, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H., and Mustapha, A. (2012). Data quality: A survey of data quality dimensions. In *2012 International Conference on Information Retrieval Knowledge Management*, pages 300–304.

Silvola, R., Harkonen, J., Vilppola, O., Kropsu-Vehkaperä, H., and Haapasalo, H. (2016). Data quality assessment and improvement. *International Journal of Business Information Systems*, 22:62–81.