AtoZ
novas práticas em informação e conhecimento

ISSN: 2237-826X
ARTIGO | PAPER

Melhoria na qualidade de dados com a aplicação de "data cleaning" na base de dados de acidentes aeronáuticos da aviação civil brasileira

Application of data cleaning to improve data quality at the Brazilian Civil Aviation Aircraft Accidents Database

Cleibson Aparecido de Almeida[1], Leonardo Derckan Rodrigues Silva[2], Elaine Cristina da Silva Schilipack[3], Nivaldo Aparecido Minervi[3]

[1]*Universidade Aberta de Portugal – UAb, Lisboa, Portugal*
[2]*Instituto de Tecnologia da Aeronáutica – ITA, São José dos Campos, SP, Brasil*
[3]*Universidade Federal do Paraná – UFPR, Curitiba, PR, Brasil*

Autor para correspondência/*Mail to:* Cleibson Aparecido de Almeida (contato@cleibsonalmeida.blog.br)

Financiamento/*Funding:* Centro de Investigação e Prevenção de Acidentes Aeronáuticos (CENIPA/FAB)

Recebido/*Submitted*: 21 Jun. 2016; **Aceito/*Approved*:** 26 Ago. 2016

Resumo

Introdução: Apresenta a aplicação de técnicas de *data cleaning* na base de dados de acidentes aeronáuticos da aviação civil brasileira com o objetivo de mensurar o grau de melhoria na qualidade dos dados.
Método: inicialmente realizou-se uma revisão de literatura sobre os conceitos de *data cleaning* e qualidade de dados e, em seguida, aplicaram-se as técnicas de *data cleaning* em uma base de dados composta por 4601 registros, referentes aos acidentes aeronáuticos ocorridos entre os anos de 1979 e 2014 na aviação civil brasileira. A medição da melhoria na qualidade dos dados foi realizada por meio da métrica "percentual de melhoria dos dados".
Resultados: Observando-se o contexto geral todos os atributos da base de dados houve uma melhoria de 9% quanto à qualidade dos dados, com atributos, como por exemplo o peso, fabricante e modelo das aeronaves, que apresentaram um grau de melhoria acima de 55% após a aplicação da metodologia.
Conclusão: A técnica de *data cleaning* pode ser utilizada para definir políticas para a melhoria contínua em bases de dados e melhorar os processos de decisão nas organizações que tratam sobre aviação, em especial na área de segurança de voo.

Palavras-chave: Limpeza de dados; Qualidade de dados; Métodos de limpeza de dados

Abstract

***Introduction**: It shows the application of techniques of data cleaning in the aeronautical accidents of brazilian civil aviation with the aim of measuring the degree of improvement in the quality of the data.*
***Method**: Initially, there was a literature review on the concepts of data cleaning and data quality, and then applied the techniques of data cleaning in a database composed of 4601 records, relating to aviation accidents that occurred between the years of 1979 and 2014 in brazilian civil aviation. The measurement of the improvement in the quality of the data was performed using the metric "percent of improvement of data".*
***Results**: Observing the general context all the attributes of the database there was a 9% improvement on the quality of the data, with attributes, such as weight, manufacturer and model of the aircraft, which had a degree of improvement over 55% after application of the methodology.*
***Conclusion**: The data cleaning technique can be used to define policies for continuous improvement in data bases and improve decision-making processes in organizations that deal with aviation, particularly in the area of flight safety.*

***Keywords:** Data cleansing; Data quality; Data cleansing methods*

INTRODUÇÃO

Nos últimos 50 anos as organizações vêm transformando seus processos baseados na economia industrial para uma economia de informação. Desde então, o mundo tem visto uma transformação da competição que se baseava na excelência produtiva, marketing e distribuição de produtos para uma atividade baseada na captação e processamento de dados como pré-requisito para definir estratégias de produção, marketing e distribuição de produtos e serviços.

Como fator fundamental para essa transformação, a popularização dos computadores, que vem ocorrendo desde meados dos anos 1990 e, mais recentemente, os dispositivos móveis, têm sido utilizados pelas empresas como o principal meio para a captura, armazenamento, visualização e descoberta de oportunidades baseadas nas informações produzidas.

Lopes (2006) ressalta que:

> Se a informação é a moeda da nova economia, então os dados são a matéria prima essencial necessária para alcançar o sucesso. Eles são os bens que formam a base dos planos estratégicos e ações que determinam o bom desempenho de um empreendimento.

Acompanhando esta tendência, a aviação civil brasileira tem coletado e armazenado diversas informações. São informações sobre rotas e horários de voos, disponibilidade de aeroportos, registros de aeronaves e tripulantes, dentre outras.

No contexto da aviação civil brasileira, a área de segurança de voo tem sido alvo de interesse da sociedade, principalmente após os dois últimos acidentes aeronáuticos que causaram grande comoção nacional – o A-022/CENIPA/2008 (Centro de Investigação e Prevenção de Acidentes [CENIPA], 2008) e o A-Nº67/CENIPA/2009 (Centro de Investigação e Prevenção de Acidentes [CENIPA], 2009), nos quais morreram 154 e 199 pessoas, respectivamente.

São apenas dois casos entre os vários acidentes aeronáuticos que ocorrem anualmente no Brasil. Somente no ano de 2014 foram registrados 145 acidentes com um total de 70 fatalidades.

De fato, os acidentes aeronáuticos acontecem em diversas condições e são causados por diferentes fatores, porém uma base de dados histórica poderia auxiliar nos trabalhos de identificação dos riscos, prevenção, redução e um melhor entendimento sobre esses acidentes.

Entretanto nem sempre uma base de dados histórica apresenta condições de qualidade para utilização em processos decisórios, visto que podem estar cercadas de problemas concernentes a anomalias de dados. Estes problemas podem ser provenientes da falta de padrão/validação na entrada de dados; da agregação de sistemas legados; da ruptura na coleta de informação em determinados períodos; entre outros.

De acordo com Kanki e Seamster (2002), uma aviação segura depende de padronizações operacionais seguindo taxonomias internacionais e isso deveria garantir uma boa qualidade nos dados gerados nas atividades aeronáuticas.

Desta forma, este trabalho tem como objetivo mensurar a melhoria na qualidade dos dados após a aplicação de um *data cleaning* (processo de limpeza de dados) na base de dados de acidentes aeronáuticos ocorridos na aviação civil brasileira entre 1979 e 2014.

Para isso, este documento está estruturado de forma a apresentar brevemente os conceitos sobre *data cleaning* e qualidade de dados, os procedimentos metodológicos adotados, a apresentação e análise dos resultados e conclusão.

REFERENCIAL TEÓRICO

Data cleaning

Data cleaning, ou limpeza de dados - também conhecida como *data cleansing* ou *scrubbing* - é um conjunto de técnicas utilizadas para detectar e remover anomalias em bases de dados (Rahm & Do, 2000).

De acordo com Vasco (2013), as anomalias existentes em bancos de dados podem ser divididas em três categorias: a) anomalias de sintaxe; b) anomalias de semântica e; c) anomalias de cobertura.

As anomalias de sintaxe condizem ao formato e valores adotados para a representação do dado. Podem conter erros lexicais, erros no formato do domínio e irregularidades. As anomalias de semântica condizem ao não entendimento do dado registrado. Podem conter violações das restrições de integridade, contradições, duplicidade de registros e registros inválidos. As anomalias de cobertura condizem à ausência de informação quando esta for uma premissa do dado. Podem conter valores omissos e registros omissos (o Quadro 1 apresenta alguns exemplos de anomalias em bases de dados).

Quanto às técnicas utilizadas para a limpeza de dados, Oliveira, Rodrigues, e Henriques (2004) destacam que elas podem ser divididas em abordagens especializadas e abordagens genéricas.

As abordagens especializadas são utilizadas sobre um determinado campo de domínio da base de dados, ou seja, foca em um problema concreto já conhecido como correção de nomes, endereços, registros duplicados.

As abordagens genéricas são utilizadas de forma sistemática na base de dados, ou seja, cobrem um campo mais vasto de problemas e adaptam-se a diferentes domínios que sejam de desconhecimento do analista. Tanto a abordagem específica, quanto a abordagem genérica, depende de técnicas (algoritmos/regras) que automatizem, de forma total ou parcial, as tarefas de detecção e correção das anomalias (Rahm & Do, 2000).

De acordo com Lopes (2006) e Vasco (2013), algumas técnicas gerais que podem ser utilizadas para a detecção de erros em *data cleaning* são: a) técnicas estatísticas; b) clusterização; c) baseado em padrão; d) regras associativas; e) análise sintática (*Parsing*); f) transformação de dados; g) aplicação de restrições de integridade e; h) eliminação de duplicatas.

Apesar do *data cleaning* ter como objetivo uma base de dados isenta de erros, isso nem sempre é alcançado. Por isso, é salutar a adoção de medidas preventivas que visem evitar futuros trabalhos com dados anômalos. Além de ser mais eficaz, a prevenção tem um custo menor do que futuras correções.

Atributo	Valor registrado	Valor esperado	Tipo de anomalia
gênero	40	'Masculino' ou 'Feminino'	Erro Lexical
localidade	São Paulo BR	'São Paulo - BR'	Erro no Formato do Domínio
gênero	Masculino | M	'Masculino' | 'Masculino'	Irregularidade
gênero	Homossexual	'Masculino' ou 'Feminino'	Violação da restrição de integridade
Gênero – gravidez?	M – SIM	'F - SIM' ou 'M-NÃO'	Contradição
localidade	São Paulo – BR | S. Paulo - BR	'São Paulo – BR' | 'São Paulo – BR'	Duplicidade de Registros
gênero	NULL	'Masculino' ou 'Feminino'	Valor Omisso
Id – gênero - localidade	23 – NULL - NULL	'23 - Masculino - São Paulo'	Registro Omisso

Quadro 1. Exemplos de anomalias de dados.
Fonte: Adaptado de Vasco (2013).

Qualidade de dados

Com foco na menor quantidade possível de anomalias, a qualidade de dados é um fator importante para determinar a precisão da tomada de decisão quando se utiliza informações provenientes de bancos de dados.

De acordo com Orr (1998), "nenhum sistema de informação tem uma qualidade dados de 100%". No entanto, a principal preocupação com a qualidade de dados não é a garantia de perfeição, mas garantir um nível de qualidade em que seja possível tomar decisões razoáveis de forma rápida e que garanta a sobrevivência de uma organização.

O nível de qualidade de um banco de dados pode ser medido de forma específica, utilizando os conceitos apresentados no trabalho de Strong, Lee, e Wang (1997). Em sua pesquisa, os autores definiram categorias e dimensões para a qualidade de dados (Quadro 2).

Categoria	Dimensão	Definição
Intrínseca	Acuracidade	A informação é correta e confiável?
	Reputabilidade	A informação espelha a sua fonte ou conteúdo?
	Credibilidade	A informação é verdadeira?
	Objetividade	A informação é imparcial?
Acessibilidade	Acessibilidade	A informação está disponível de forma fácil?
	Segurança	Até quanto a informação é restrita para resguardar a sua segurança?
Contextualidade	Relevância	A informação é útil para a tarefa em questão?
	Completude	A informação é completa e suficiente ao nível de largura e profundidade?
	Temporalidade	A informação está atualizada para a tarefa em questão?
	Valor agregado	A informação é benéfica e há vantagens em sua utilização?
	Quantidade	O volume de informação é adequado para a tarefa em questão?
Representatividade	Interpretação	A informação está em linguagem, símbolo e unidade apropriada para a tarefa em questão? Sua definição é clara?
	Facilidade de uso	A informação é facilmente manipulada e aplicada em diferentes tarefas?
	Facilidade de entendimento	A informação é facilmente entendida?
	Representação concisa	A informação está representada de forma compacta?
	Representação consistente	A informação está representada no mesmo formato para um mesmo atributo em diferentes registros?

Quadro 2. Categorias e dimensões na qualidade de dados.
Fonte: Adaptado de Strong et al. (1997) e Pipino, Lee, e Wang (2002).

Outra forma específica para medir a qualidade dos dados foi descrita por Vasco (2013), conforme os itens a seguir: a) precisão: é o quociente entre o número de valores corretos e o número total de valores existentes no conjunto de dados; b) integridade: este critério é satisfeito quando não existem registros inválidos, nem violações de restrição de integridade ou valores omissos; c) completude: é o percentual de preenchimento em um registro, considerando que todos os atributos considerados de preenchimento obrigatório devam estar alimentados; d) validade: é o percentual de registros que são entidades válidas. Registros que violem as condições de integridade são considerados inválidos; e) consistência: é o percentual de registros que possuem anomalias de sintaxe ou contradições; f) conformidade de esquema: é o percentual registros que estão conforme a estrutura sintática definida pelo esquema relacional; g) uniformidade: é o percentual de registros que não contém irregularidades nos seus valores; h) densidade: é o percentual de valores não omissos nos registros e; i) unicidade: percentual de registros únicos, ou seja, registros sem duplicidades.

MÉTODO

Descrição do conjunto de dados

O conjunto de dados utilizado na pesquisa corresponde ao número total de acidentes aeronáuticos ocorridos com aeronaves brasileiras, entre 1979 e 2014, resultando em 4601 registros (eventos).

Foram utilizadas as duas principais entidades do banco de dados, com um total de 19 (dezenove) atributos, conforme ilustrado na Figura 1.

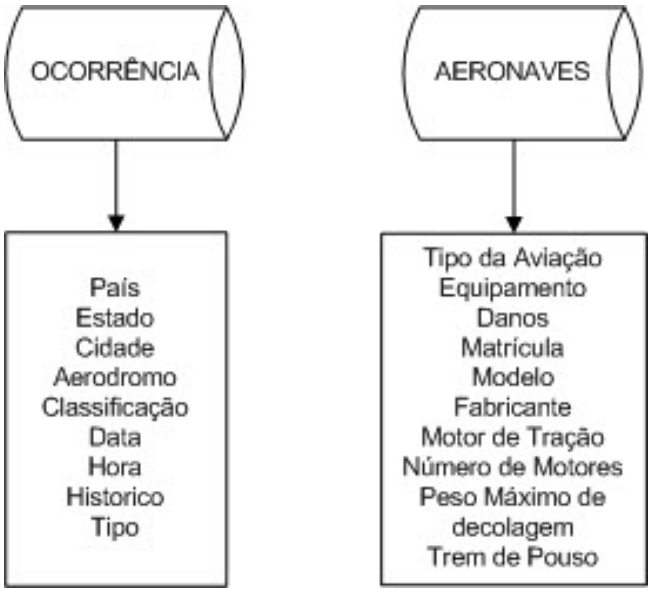

Figura 1. Entidades Ocorrência e Aeronaves.
Fonte: Elaborado pelos autores.

Detalhamento dos atributos

O Quadro 3 apresenta a descrição e exemplos para cada um dos 19 atributos utilizados no estudo.

Aplicação do Data Cleaning no banco de dados

A aplicação do *data cleaning* foi realizada mediante um processo controlado e mensurado após a atualização do banco de dados (Figura 2).

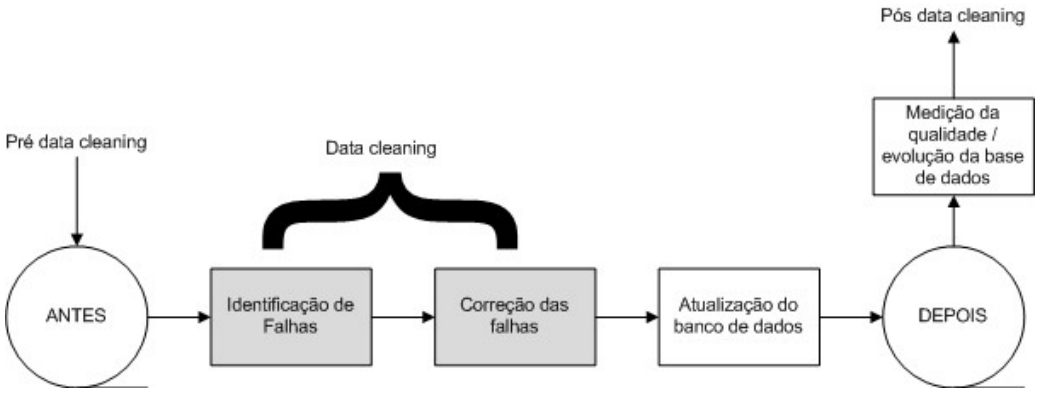

Figura 2. Processo para aplicação do *data cleaning*.
Fonte: Elaborado pelos autores.

Utilizando a abordagem especializada, o *data cleaning* foi voltado à erros lexicais, erros no formato de domínio, irregularidades nos dados, violações de restrição de integridade, duplicidade de registros, valores omissos e registros omissos.

Para isso foram adotadas as técnicas estatísticas para correção dos atributos numéricos, baseado em padrão. Nos atributos do tipo data, hora e texto foram aplicadas regras associativas.

A execução das atividades foi parcialmente automatizada com o apoio dos seguintes softwares: a) QLIKVIEW 11; b) MS EXCEL 2007 e, c) R PROJECT 3.0.

Atributo (tipo de dado)	Descrição	Exemplo
País (texto)	Nome do país onde ocorreu o acidente.	'BRASIL', 'PERU', 'CHILE'...
Estado (texto)	Sigla do Estado/Província onde ocorreu o acidente.	'SP', 'EX', 'EX'...
Cidade (texto)	Nome da cidade onde ocorreu o acidente.	'SÃO PAULO', 'LIMA', 'SANTIAGO'...
Aeródromo (texto)	Código ICAO* da pista de pouso/decolagem onde ocorreu o acidente (composto por 4 caracteres).	'SBGR', 'SBTY', 'KLOP'...
Classificação (texto)	Classificação da ocorrência. Neste caso apenas acidentes.	'ACIDENTE'
Data (data)	Data do acidente no formato dia/mês/ano no padrão numérico.	'01/09/2011', '15/02/1979'...
Hora (hora)	Horário local do acidente no formato hora:minuto no padrão 24h.	'11:20', '18:59', 23:45'...
Histórico (texto)	Texto resumo que descreve a ocorrência. Pode ser qualquer conjunto de frases formadas por qualquer tipo de caracteres.	'A aeronave PTRFG estava sobrevoando o aeroporto SBGR quando colidiu com um urubu'...
Tipo (texto)	Taxonomia de tipo de ocorrência aeronáutica prevista no MCA 3-6 .	'COLISÃO COM OBSTÁCULO NO SOLO'...
Tipo da Aviação (texto)	Define se a aeronave é CIVIL ou MILITAR. Nesta pesquisa será considerada apenas a CIVIL.	'CIVIL'
Equipamento (texto)	Tipo de aeronave que se acidentou. Pode ser um AVIÃO, HELICÓPTERO ou outro conforme o RAB**.	'AVIÃO', 'PLANADOR', 'HELICÓPTERO'...
Danos (texto)	Nível do dano da aeronave naquele acidente. Pode assumir um entre quatro diferentes valores.	'NENHUM', 'LEVE', 'SUBSTANCIAL', 'DESTRUÍDA'
Matrícula (texto)	Marca da aeronave registrada no RAB**. É composta por 5 ou mais caracteres.	'PTOLK', 'PUGHT', 'N600XL'...
Modelo (texto)	Modelo de fabricação da aeronave.	'BOEING 757', 'EMB 190'...
Fabricante (texto)	Nome do fabricante da aeronave.	'EMBRAER', 'BOEING', 'BOMBARDIER'...
Motor de Tração (texto)	Tipo do motor da aeronave.	'JATO', 'TURBOÉLICE'...
Número de Motores (número)	Quantidade de motores presente na aeronave. Podendo ser 0 ou mais.	'0', '1'...'4'...
Peso Máximo de Decolagem (número)	Peso máximo de decolagem de uma aeronave.	'1263', '123890'...
Trem de Pouso (texto)	Tipo do trem de pouso da aeronave.	'CONVENCIONAL', 'TRICICLO'...

Quadro 3. Descrição dos atributos utilizados na pesquisa.
Fonte: Elaborado pelos autores.
Notas: *ICAO: International Civil Aviation Organization. **RAB: Registro Aeronáutico Brasileiro (http://www2.anac.gov.br/aeronaves/cons_rab.asp).

O QLIKVIEW foi utilizado para fazer a extração dos dados disponíveis no Sistema Gerenciador de Banco de Dados (SGBD), em ORACLE 10G DATABASE, da Organização. O MS EXCEL foi utilizado como ferramenta auxiliar nas atividades de visualização dos dados em formato tabular e preparação de dados para reinserção no SGBD após o *data cleaning*. O R-PROJECT foi utilizado para identificação e correção das falhas identificadas.

Métrica utilizada para a qualidade dos dados

A métrica utilizada para a medição da qualidade dos dados foi a do percentual de melhoria dos dados, ou seja, $(QEA \div N) \times 100$, sendo que o QEA representa a quantidade de erros encontrados na situação pré *data cleaning* em comparação com a situação pós *data cleaning* e N representa o total de registros na base dados.

Para exibição dos resultados foram utilizados gráficos estatísticos apresentando o percentual de melhoria dos dados em cada um dos atributos considerados no estudo.

RESULTADOS E DISCUSSÃO

Os resultados apresentados nesta seção mostram a melhoria da base de dados após aplicação do *data cleaning* quanto às seguintes anomalias: a) correção de erros lexicais; b) correção de erros de formato no domínio; c) eliminação de irregularidades; d) melhoria nas restrições de integridade; e) eliminação de registros duplicados; f) preenchimento de valores omissos e g) eliminação de registros omissos.

Em relação à melhoria dos dados da entidade Ocorrência, pode-se observar na Figura 3 o percentual de melhoria em cada um dos atributos que passaram pelo *data cleaning*. De forma geral, o erro médio encontrado nesta entidade foi de 6% (seis por cento).

Os atributos Aeródromo e Cidade apresentaram os maiores percentuais de erros, ou seja, 49% e 38.4%, respectivamente. Dentre as anomalias identificadas nestes dois atributos, concentraram-se esforços para melhorar as restrições de integridade, corrigir erros no formato do domínio e o preenchimento de valores omissos.

Os demais atributos da entidade Ocorrência apresentaram um percentual abaixo do erro médio (6%), sendo que o atributo País teve um percentual de melhoria igual a 0, ou seja, não continha erros antes da aplicação do *data cleaning*.

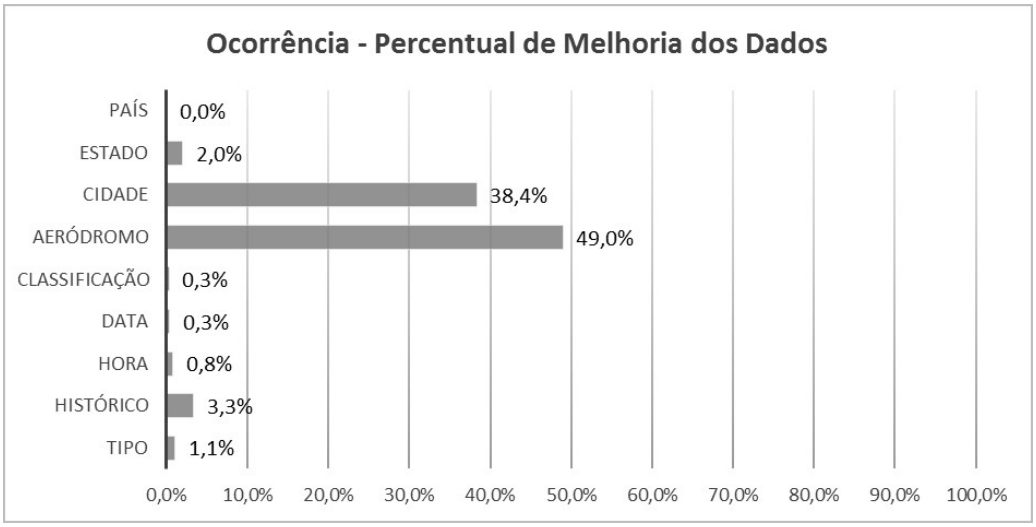

Figura 3. Percentual de melhoria na entidade Ocorrência – Base de dados de acidentes aeronáuticos da aviação civil brasileira.
Fonte: Elaborado pelos autores, via análise dos dados.

Quanto à entidade Aeronave, o percentual de erros encontrados e corrigidos foi maior quando comparado com a entidade Ocorrência. Enquanto a Ocorrência apresentou um erro médio de 6%, a entidade Aeronave teve uma média de erro igual a 12%.

A Figura 4 apresenta o percentual de melhoria realizada na entidade aeronave. Os atributos que apresentaram maior incidência de erros foram o Peso Máximo de Decolagem (98.4%), Trem de Pouso (84%), Modelo (77.2%) e Fabricante (56.1%).

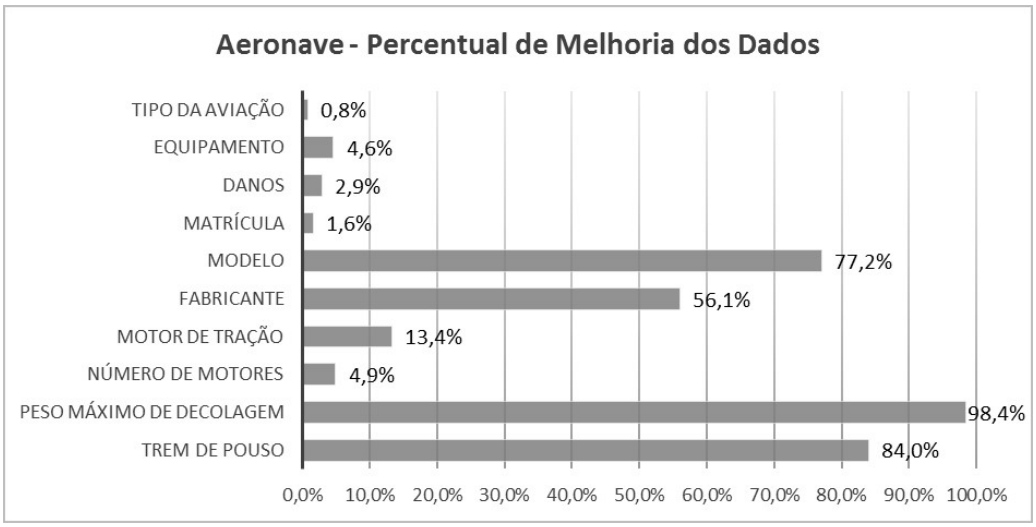

Figura 4. Percentual de melhoria na entidade Aeronave - Base de dados de acidentes aeronáuticos da aviação civil brasileira.
Fonte: Elaborado pelos autores, via análise dos dados.

Uma simples analogia com os valores brutos dessa amostra levaria a considerar que esta base de dados, composta por 4601 acidentes aeronáuticos, tinha mais de 4500 aeronaves com problemas relativos a informação do Peso Máximo de Decolagem.

Com isso, é possível inferir que a tomada de decisão com esses dados antes da aplicação do *data cleaning*, aqui proposto, proporcionava um erro de decisão de 98.4% ao gestor que utilizava esta informação. Este valor significativo quanto ao erro de decisão no Peso Máximo de Decolagem revelou três principais problemas existentes no gerenciamento no SGBD, como:

a) por ser um atributo de preenchimento não obrigatório no SGBD, na maioria das vezes o usuário não se atentava com preenchimento desta informação no banco de dados. Conforme literatura apresentada, isso gerou erros do tipo "valor omisso";

b) por ser uma informação proveniente de terceiros, o usuário necessitava consultar e confiar na credibilidade de uma base de dados externa, o que nem sempre garantia que a informação foi corretamente consultada pelo usuário. Conforme literatura apresentada, isso gerou "erros no formato do domínio";

c) por requerer um grau de conhecimento *a priori* quanto aos modelos de aeronaves existentes, muitas vezes

o usuário preenchia este campo sem realizar uma pesquisa mais aprofundada sobre o assunto. Com isso, o campo era preenchido com valores baseados em conhecimentos prejulgados pelos usuários, baseados em aeronaves similares e distorcidos da realidade especifica daquela aeronave. Conforme literatura apresentada, isso gerou erros do tipo "violação da restrição de integridade".

Estes três tipos de anomalias também foram identificadas em outros três atributos (Trem de pouso, Modelo e Fabricante), que apresentaram um percentual de erros em 84%, 77.2% e 56.1%, respectivamente.

Em contrapartida, os atributos Tipo da Aviação, Equipamento, Danos, Matrícula e Número de Motores apresentaram um percentual de melhoria abaixo do erro médio desta entidade (12%). Os erros encontrados nesses atributos estavam ligados a duplicidade de registros e erros no formato do domínio.

No contexto geral, a melhoria da base de dados utilizado neste estudo foi de 9%. Isso quer dizer que a cada 100 acidentes aeronáuticos contidos nesta base de dados, 9 (nove) possuíam erros que poderiam afetar uma tomada de decisão.

Em consulta verbal aos especialistas que trabalham há mais de dez anos com essa base de dados, foi percebido que a integração ineficiente de sistemas legados, sistemas com falhas na entrada de dados, processos operacionais instáveis e a falta de comprometimento com a qualidade dos dados foram alguns dos fatores que corroboraram para os problemas de qualidade dos dados utilizados nesta pesquisa. Este fato vai ao encontro do proposto na pesquisa realizada por Oliveira et al. (2004) e ao mesmo tempo comprova a afirmação de Orr (1998), sobre a dificuldade em obter uma qualidade de dados 100% em um sistema de informação.

CONCLUSÃO

O presente estudo apresentou a aplicação da técnica de *data cleaning* em uma base de dados de acidentes aeronáuticos ocorridos entre 1979 e 2014, no Brasil. Nesse cenário, vale ressaltar que os especialistas que trabalham há mais de dez anos com a base de dados utilizada nesta pesquisa afirmaram que a utilização de sistemas com falhas na entrada dos dados, problemas com integração de sistemas legados, instabilidade nos processos operacionais e a falta de comprometimento com a qualidade dos dados foram os fatores causadores da baixa qualidade dos dados.

Dentre as anomalias identificadas na base de dados utilizada, destaca-se que o preenchimento de atributos com valores omissos e correções de erros no formato dos domínios e violações nas restrições de integridade são aquelas que mais influenciaram na melhoria da qualidade dos dados.

Com isso, sugere-se que antes de iniciar algum procedimento de *data cleaning* em uma base de dados, sejam observadas a existência desses três tipos de anomalias, visto que este pequeno esforço inicial poderá melhorar significativamente a qualidade dos dados.

Desta forma, a pesquisa atingiu o seu objetivo, ou seja, mensurar o grau de melhoria na qualidade dos dados após a aplicação da técnica de *data cleaning* na base de dados de acidentes aeronáuticos da aviação civil brasileira. Fica conclusivo que a aplicação da técnica de *data cleaning* é uma boa prática para a definição de políticas que foquem na melhoria contínua em bases de dados utilizadas para a tomada de decisão, neste caso específico, auxiliar no processo decisório durante a identificação de riscos na prevenção de acidentes aeronáuticos.

Os autores propõem que se realizem pesquisas adicionais em áreas fora do escopo da aviação civil brasileira, para identificar limitações ou lacunas onde a técnica de *data cleaning* seja relevante para a melhoria da qualidade dos dados. Esta proposta também ajudará na generalização dos resultados até aqui alcançados.

REFERÊNCIAS

Centro de Investigação e Prevenção de Acidentes. (2008). *Relatório final a-022/cenipa/2008*.

Centro de Investigação e Prevenção de Acidentes. (2009). *Relatório final a-nº67/cenipa/2009*.

Kanki, B. G., & Seamster, T. L. (2002). *Aviation information management: From documents to data*. Burlington: Ashgate.

Lopes, F. P. (2006). *Administração de dados: Técnicas, metodologias e ferramentas para garantir a qualidade dos dados*. Recife: Universidade Federal de Pernambuco.

Oliveira, P. J., Rodrigues, F., & Henriques, P. R. (2004). *Limpeza de dados: Uma visão geral.* Recuperado de http://wiki.di.uminho.pt/twiki/pub/Research/Doutoramentos/SDDI2004/ArtigoOliveira.pdf

Orr, K. (1998, Feb.). Data quality and systems theory. *Communications of the ACM*, *41*(2), 66–71. doi: 10.1145/269012.269023

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002, Apr.). Data quality assessment. *Communications of the ACM*, *45*(4), 211–218. doi: 10.1145/505248.506010

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, *23*(4), 3–13. Recuperado de http://sites.computer.org/debull/A00dec/issue1.htm

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997, May). Data quality in context. *Communications of the ACM*, *40*(5), 103–110. doi: 10.1145/253769.253804

Vasco, D. O. (2013). *Identificação de anomalias contextuais*. Porto: Universidade do Porto.

Como citar este artigo (APA):
Almeida, C. A., Silva, L. D. R., Schilipack, E. C. S. & Minervi, N. A. (2016). Melhoria na qualidade de dados com a aplicação de "data cleaning" na base de dados de acidentes aeronáuticos da aviação civil brasileira. *AtoZ: novas práticas em informação e conhecimento, 5*(2), 72 – 79. Recuperado de: http://dx.doi.org/10.5380/atoz.v5i2.47303