

AVALIAÇÃO DA QUALIDADE DOS DADOS DO PROCESSO DE CONSUMO E ABASTECIMENTO DE COMBUSTÍVEL DE UMA FROTA DE ÔNIBUS SOB A DIMENSÃO DA PRECISÃO

EVALUATION OF THE DATA QUALITY OF THE FUEL SUPPLY AND CONSUMPTION PROCESS OF A BUS FLEET UNDER THE DIMENSION OF PRECISION

EVALUACIÓN DE LA CALIDAD DE LOS DATOS DEL PROCESO DE SUMINISTRO Y CONSUMO DE COMBUSTIBLE DE UNA FLOTA DE AUTOBUSES BAJO LA DIMENSIÓN DE PRECISIÓN

Anderson Oberdan Machado

<https://orcid.org/0000-0002-7542-8226>

Mestrando em Gestão de Organizações, Liderança e Decisão. (PPGOLD),UFPA

Egon Walter Wildauer

<https://orcid.org/0000-0003-2340-8984>

Doutorado em Engenharia Florestal. Coordenador MBA Gestão em Engenharia - UFPR

Editor Científico: José Edson Lara
Organização Comitê Científico
Double Blind Review pelo SEER/OJS
Recebido em 10/10/2021
Aprovado em 08/07/2022

This work is licensed under a Creative Commons Attribution – Non-Commercial 3.0 Brazil

Resumo

Objetivo do estudo: Este trabalho mede a qualidade dos dados gerados no processo de consumo e abastecimento de combustível e associados na composição dos custos, com uso de informações de empresas de transporte coletivo da região de Curitiba-PR.

Metodologia/abordagem: A partir do mapeamento das atividades e dos dados coletados no processo, foram definidas e aplicadas regras de negócio e técnica estatística que permitiram calcular o percentual de precisão dos dados.

Originalidade/Relevância: A lacuna teórica se posiciona na necessidade de definir e testar métodos que qualifiquem os dados utilizados nos processos decisórios, algo ainda incipiente no meio acadêmico/organizacional.

Principais resultados: A aplicação do método demonstrou que o processo real é representado com 87,36% de precisão por meio dos seus dados.

Contribuições teóricas/metodológicas: Apresenta-se como contribuição uma metodologia de avaliação de qualidade de dados, com base na dimensão da precisão, por meio da localização de informações que não expressam fielmente o processo real.

Contribuições sociais / para a gestão: A principal contribuição gerencial é a melhoria da acuracidade das decisões a partir da consciência da precisão dos dados coletados no processo, assim como, a motivação para a criação de um programa formal de gerenciamento de qualidade dos dados.

Palavras-chave: Qualidade de dados; consumo de combustível; transporte público; dimensão da precisão.

Abstract

Objective of the study: This work measures the quality of the data generated in the fuel consumption and supply process and associated with the composition of costs, using information from public transportation companies in the Curitiba-PR region.

Methodology / approach: Based on the mapping of activities and data collected in the process, business rules and statistical techniques were defined and applied that allowed the percentage of data accuracy to be calculated.

Originality / Relevance: The theoretical gap is positioned in the need to define and test methods that qualify the data used in decision-making processes, something that is still incipient in the academic / organizational environment.

Main results: The application of the method demonstrated that the real process is represented with 87.36% accuracy through its data.

Theoretical / methodological contributions: As a contribution, a data quality assessment methodology is presented, based on the dimension of precision, by locating information that does not accurately express the real process.

Social / management contributions: The main managerial contribution is to improve the accuracy of decisions based on an awareness of the accuracy of the data collected in the process, as well as the motivation for creating a formal data quality management program.

Keywords: Data quality; fuel consumption; public transportation; precision dimension.

Resumen

Objetivo del estudio: Este trabajo mide la calidad de los datos generados en el proceso de consumo y suministro de combustible y asociados a la composición de costos, utilizando información de empresas de transporte público de la región de Curitiba-PR.

Metodología / enfoque: A partir del mapeo de actividades y datos recogidos en el proceso, se definieron y aplicaron reglas de negocio y técnicas estadísticas que permitieron calcular el porcentaje de precisión de los datos.

Originalidad / Relevancia: La brecha teórica se posiciona en la necesidad de definir y probar métodos que cualifiquen los datos utilizados en los procesos de toma de decisiones, algo que aún es incipiente en el ámbito académico / organizacional.

Resultados principales: La aplicación del método demostró que el proceso real se representa con un 87,36% de precisión a través de sus datos.

Aportes teórico-metodológicos: Como contribución, se presenta una metodología de evaluación de la calidad de los datos, basada en la dimensión de precisión, mediante la localización de información que no expresa con precisión el proceso real.

Contribuciones sociales / de gestión: La principal contribución de la gestión es mejorar la precisión de las decisiones basadas en el conocimiento de la precisión de los datos recopilados en el proceso, así como la motivación para crear un programa formal de gestión de la calidad de los datos.

Palabras clave: Calidad de los datos; consumo de combustible; transporte público; dimensión de precisión.

1. INTRODUÇÃO

O serviço de Transporte Público Coletivo (TPC) é um dos principais elementos de desenvolvimento das grandes cidades (NTU, 2016); composto de um conjunto de atividades onde o consumo e abastecimento de combustível tem relevância em função do impacto na composição dos custos. (NTU, 2018; URBS, 2020). Dados do processo de consumo de combustível servem de apoio à tomada de decisões, permitindo melhorias na produtividade (Brynjolfsson et al. 2011; Shankaranarayan et al. 2003); então, dados de baixa qualidade podem interferir negativamente nas decisões e resultados, pois não retratam com fidelidade o mundo real. Neste sentido, a Qualidade de Dados (QD) é conceito multidimensional que descreve a adequação dos objetos de dado para utilização dentro de um contexto (Wang & Strong, 1996); e aliar a atribuição de métricas às dimensões importantes aos consumidores de dados permite, através de heurísticas adequadas, o cálculo da precisão com que um processo é representado por meio de seus dados. (Wang et al. 2005).

O problema apresentado neste trabalho é fundamentado em atividades mapeadas e dados coletados de empresas que realizam o TPC na região de Curitiba, Paraná; a programação de linhas e horários destas empresas são variadas, sendo agrupadas de acordo com características similares de operação, distanciamento de pontos de paradas e distribuição da demanda. (Kuah; Perl, 1988). A Figura 1 apresenta detalhes dos grupos de operação onde a frota atua.

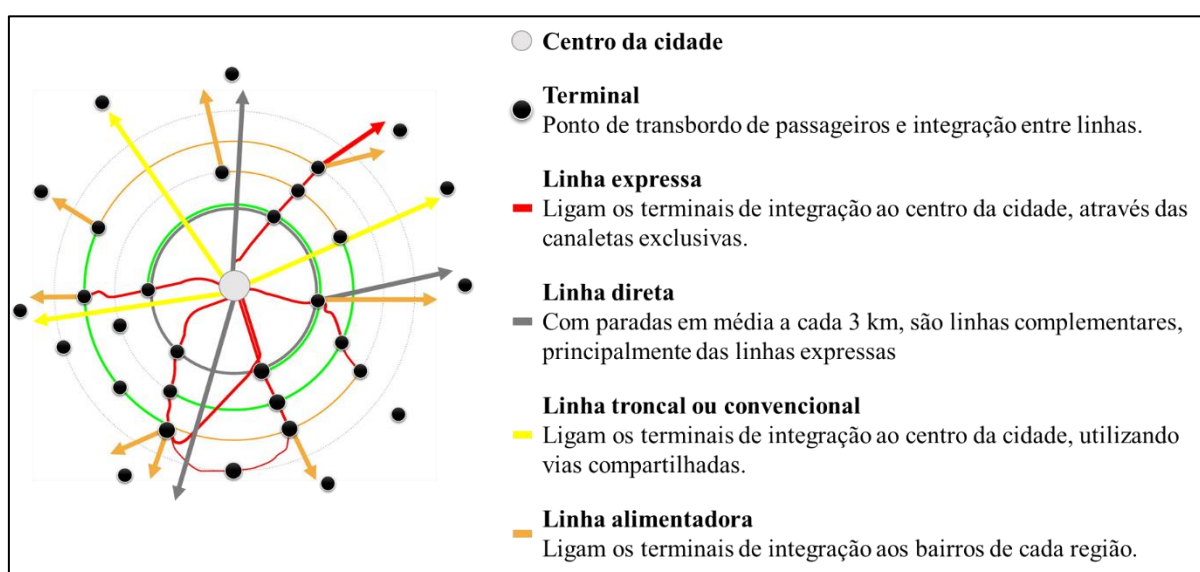





Figura 1. Estrutura dos grupos de operação de linhas de ônibus

Fonte: Adaptado de URBS. Urbanização de Curitiba S/A, 2021.

A frota de ônibus que realiza o TPC é composta por diferentes fabricantes de chassis e carrocerias as quais, segundo a Norma Técnica Brasileira (NBR) 15570:2009, são estruturas montadas sobre o chassi-plataforma adequados ao transporte de passageiros (ABNT, 2009); assim, a combinação de chassis e carrocerias resulta em ônibus de diferentes classes, detalhadas na Tabela 1:

Tabela 1
Classes e características dos veículos utilizados no TPC

Tipo	Classe	Capacidade ^a	Peso bruto	Comprimento máximo	Capacidade combustível	Máxima utilização diária
	Comum	Mínimo de 70	16	14	300	700
	Padron Semipadron	Mínimo de 80	16	14 ^b	300	700
	Articulado	Mínimo de 100	26	18,6	600	700

Nota: Capacidade medida em número de passageiros. Peso bruto medido em toneladas. Comprimento máximo medido em metros. Capacidade de combustível medida em litros. Utilização máxima diária medida em quilômetros.

Fonte: Adaptado de NBR 15570:2009 ABNT (2009).

^a Passageiros sentados e em pé, incluindo área reservada para acomodação de cadeira de rodas ou cão-guia.

^b Admite-se ônibus Padron de até 15 m, desde que o veículo seja dotado de terceiro eixo de apoio direcional.

As classes de veículo são definidas em função do tamanho, peso e capacidade de transporte de passageiros de cada equipamento, sendo designadas a grupos de operação mais adequados; a capacidade dos tanques de armazenamento de combustível é determinada pelo fabricante do chassi, enquanto que a máxima utilização diária é estabelecida de maneira empírica pelos gestores do processo, baseada no tempo médio e velocidade média de operação de cada veículo, sendo parametrizada no sistema de *Enterprise Resources Planning* (ERP).

Não raro, o mesmo veículo opera em diferentes condições de tráfego, relevo e lotação; sofre influência do peso, da dirigibilidade, etc. fazendo com que o desempenho do consumo de combustível mude de maneira considerável, requerendo ações de gestão que objetivem maximizar seu resultado.

O abastecimento de combustível acontece nas dependências da empresa, tendo seus dados registrados e controlados por meio de software de ERP; a dinâmica de atividades deste processo está representada na Figura 2, destacando em amarelo as atividades que geram e tratam os dados no processo.

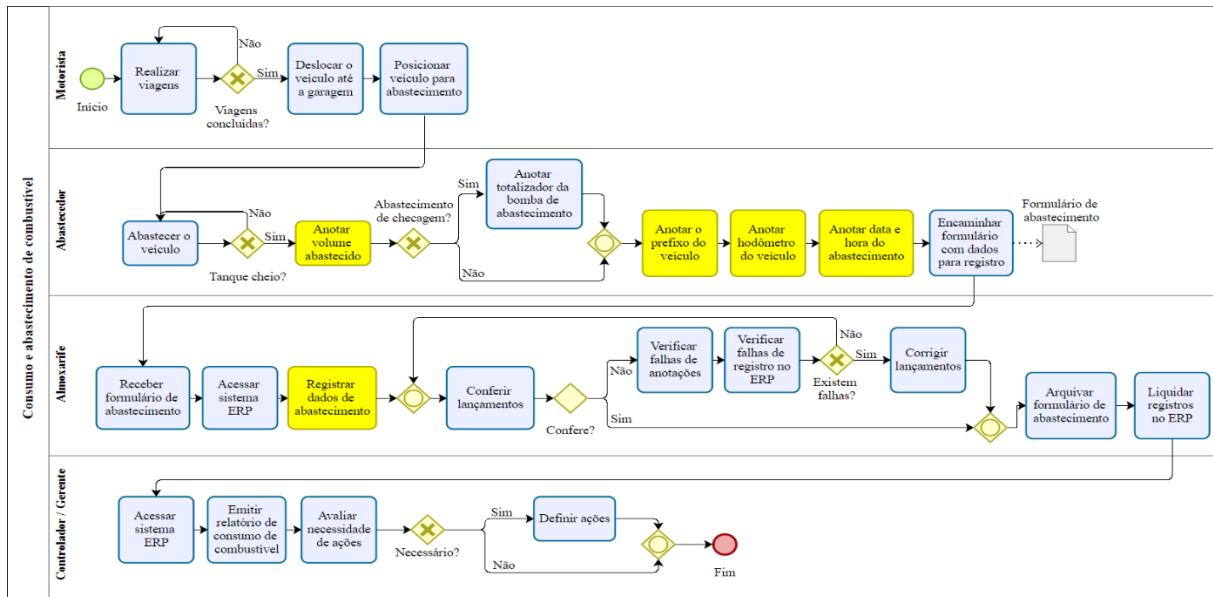


Figura 2: Processo de consumo e gestão de combustível e abastecimento dos veículos
Fonte: Elaborado pelos autores.

Ao concluir a operação diária, o motorista desloca o veículo até a garagem da empresa, para que um funcionário realize o abastecimento preenchendo o tanque de combustível completamente; com isso, é constatada a quantidade de combustível consumida durante o dia. O funcionário então anota em um formulário os seguintes dados: prefixo do veículo; data e hora do abastecimento; hodômetro (quilometragem) do veículo; volume (em litros) abastecidos e, quando necessário, o numerador da bomba de abastecimento, encaminhando o formulário para registro no sistema de ERP.

Após cadastramento dos dados, são emitidas listagens que apresentam o desempenho de consumo de combustível em diferentes períodos, sendo a avaliação do consumo médio por quilômetro, destacado na Figura 3, a principal informação utilizada pelos gestores do processo para definição de ações.

Equipamento	Registrador		Utilização	Qtde	Média	ΔM%	M.I.	Qtde M.I.	Diferença	ΔMI%
001	374.662	4	1.629	706,1	2,3070	-14,31	2,0000	815	-108	15,35
002	368.542	4	1.300	430,0	3,0233	12,29	2,0000	650	-220	51,16
004	363.249	3	528	186,0	2,8387	5,43	2,0000	264	-78	41,94
005	252.021	2	1.380	441,9	3,1229	15,99	2,0000	690	-248	56,14
006	264.612	2	1.783	613,8	2,9049	7,89	2,0000	892	-278	45,24
007	272.745	2	2.281	746,4	3,0560	13,50	2,0000	1.141	-394	52,80
008	226.208	2	1.675	591,2	2,8332	5,23	2,0000	838	-246	41,66
009	219.500	1	2.168	708,0	3,0621	13,73	2,0000	1.084	-376	53,11
010	197.412	1	2.413	772,2	3,1248	16,06	2,0000	1.207	-434	56,24
011	122.452	1	2.123	773,9	2,7432	1,89	2,0000	1.062	-288	37,16
012	116.435	1	1.066	390,1	2,7326	1,49	2,0000	533	-143	36,63
013	83.799	1	2.278	751,9	3,0297	12,53	2,0000	1.139	-387	51,48
014	78.784	1	1.786	585,1	3,0525	13,37	2,0000	893	-308	52,62
015	57.545	0	1.295	468,2	2,7659	2,73	2,0000	648	-179	38,30
016	55.603	0	1.546	520,5	2,9702	10,32	2,0000	773	-253	48,51

Figura 3. Relatório de média de consumo de combustível, emitido a partir do ERP

Fonte: Elaborado pelos autores.

Eventuais variações nos resultados podem ser provenientes de alterações na dinâmica de operação dos veículos, problemas de manutenção ou inconsistência nos dados; todavia, não existe um procedimento estabelecido que verifique quando os dados infringem as regras do negócio, ou estão demasiadamente distantes dos demais registros, caracterizando a imprecisão em representar o processo real no qual foram gerados.

Diante deste cenário, o presente trabalho mede, utilizando regras de negócio e técnica estatística, a precisão dos dados gerados no processo de consumo e abastecimento de combustível de empresas de TPC da região de Curitiba-PR; e que são utilizados no processo de tomada de decisões.

2. REFERENCIAL TEÓRICO

A QD é um fator crítico para atingir os objetivos estratégicos e operacionais do negócio; entre esses objetivos estão a melhoria na tomada de decisões (Shankaranarayan et al., 2003; Price e Shanks, 2005); nesta mesma linha, Choo (2003) afirma que o uso estratégico

da informação pode desenvolver uma organização do conhecimento que cria, organiza e processa informações que servem de guia para a tomadas de decisões racionais importantes; da mesma forma, Brynjolfsson et al. (2011) afirmam que a tomada de decisões orientada por dados está associada a melhoria da produtividade. Porém, funcionários passam menos tempo tentando descobrir se os dados estão corretos e mais tempo usando os dados para tomar decisões e obter insights (Earley; Henderson, 2017); com isso, os riscos da utilização de dados de baixa qualidade, que não retratam de forma consistente os processos, são ainda maiores.

Fundamentado nos motivadores de negócios, o *Data Management Body of Knowledge* (DAMA-BOK®) propõe a criação de um programa formal de gerenciamento de qualidade dos dados; proporcionando a redução nos riscos e custos associados a dados de baixa qualidade, melhoria da eficiência organizacional e da produtividade e redução de perdas devido à más decisões de negócios motivadas por dados ruins. (Earley e Henderson, 2017);

Barbieri e Farinelli (2013) afirmam que a análise das regras de negócios fundamentais dos processos, ou seja, restrições com base em descrições e fatos (Zoet et al., 2011), é necessária para a descoberta dos dados que podem implicar em quebras de conformidade; todavia, de acordo com o DAMA-BOK® os analistas precisam identificar as regras de negócios que descrevem ou implicam em expectativas sobre as características de qualidade dos dados, mas eventualmente estas regras não são documentadas explicitamente (Earley e Henderson, 2017). O mesmo recurso de dados pode ter um nível de qualidade aceitável em alguns contextos e inaceitável em outros, sendo papel dos consumidores de dados avaliar as regras de qualidade, em diferentes contextos de negócios ou dimensões específicas. (Even e Shankaranarayanan, 2007).

Earley e Henderson (2017) estabelecem dimensão como um termo usado para fazer a conexão com as dimensões na medição de objetos físicos (por exemplo, comprimento, largura, altura); o *Data Management Association (DAMA) UK Working Group on "Data Quality Dimensions"* (2013), descreve dimensão como um recurso de dados que pode ser medido ou avaliado em relação a requisitos mensuráveis, padrões e expectativas, onde a qualidade dos dados pode ser medida de forma objetiva.

De acordo com Bizer e Cyganiak (2009) as métricas empregadas na medição da qualidade dos dados, podem ser apoiadas em metainformações sobre as circunstâncias em que

os dados foram criados, em informações básicas sobre o provedor ou em classificações fornecidas pelo próprio consumidor de dados ou especialistas no domínio. Técnicas estatísticas também são empregadas no contexto dos negócios como valiosas ferramentas de medição; Anderson et al. (2016) afirmam que as estatísticas podem ser chamadas de fatos numéricos e que num sentido mais amplo, estatística é o campo de estudo que trata da coleta, análise, apresentação e interpretação de dados.

3. METODOLOGIA

Considerando que o objetivo deste trabalho é verificar o grau de precisão que os dados representam o processo real, buscou-se à partir de uma amostra, localizar pontos de dados divergentes que desviam dos demais, chamados *outliers* (Hawkins, 1980); e também dados incorretos, os quais de acordo com Aguinis et al. (2013), são um tipo de *outlier* que representam observações ilegítimas do processo; a partir destes valores, calcular então a representatividade sobre a amostra de registros coletados, obtendo assim a medida de precisão.

Foram definidas regras de negócios de qualidade de dados, as quais, descrevem como os dados devem existir para serem úteis e utilizáveis na organização; estas regras representam faixas de intervalo de valores ou averiguação de precisão, permitindo a comparação de valores próximos ou dentro de uma faixa correspondente. (Earley e Henderson 2017).

Cinco regras foram definidas, sendo quatro delas utilizadas para identificar a presença de observações ilegítimas do processo (*outlier* de erro) e uma regra utilizada para identificar dados divergentes (*outlier* divergente). A Tabela 2 detalha estas regras.

Tabela 2

Regras para identificação de dados com erros ou divergentes

Regra	Tipo de <i>Outlier</i>	Descrição
R1	Erro	O volume de combustível (LT) inserido no tanque de armazenamento é maior que a capacidade máxima do tanque do veículo (CM);
R2	Erro	A distância diária (D) percorrida pelo veículo é maior que limite máximo de operação diária (LO) definido como parâmetro no sistema ERP;
R3	Erro	O veículo realizou operação ($D > 0$), porém não houve abastecimento de combustível ($LT = 0$);
R4	Erro	O veículo foi abastecido de combustível ($LT > 0$), porém não houve registro de operação ($D = 0$);
R5	Divergente	Valor está fora do espaço contínuo de entrada identificado por meio da técnica do intervalo interquartil;

Fonte: Elaborado pelos autores.

O intervalo interquartil (IQR), utilizado para o cálculo dos *outliers* divergentes, é uma técnica estatística que ajuda a encontrar desvios em dados continuamente distribuídos; é calculado pela subtração do valor do terceiro quartil (Q3) pelo primeiro quartil (Q1) (Vinutha et al. 2018); uma instância de dados que está além dos *whiskers* (*bigodes*) ou seja, a mais que $1,5 * IQR$ menor que Q1 ou $1,5 * IQR$ maior que Q3, são declarados como anomalias (Chandola et al., 2009).

A Figura 4 apresenta o modelo conceitual proposto neste trabalho para a aplicação das regras e cálculo da precisão dos dados. Na primeira etapa, são aplicadas sob a amostra de dados do processo as regras de negócios (R1 a R4); com isso, são identificados os *outliers* de erros; os dados que não infringem as regras de negócio, são considerados válidos e a partir deles é aplicada a técnica estatística descrita em R5, obtendo assim, os *outliers* divergentes; os dados que não fazem parte do conjunto de *outliers* de erros e *outliers* divergentes, são considerados *inliers*. A soma de todos os *outliers* identificados pelo conjunto das cinco regras

é dividida pelo número de elementos da amostra, obtendo assim, o percentual de precisão dos dados.

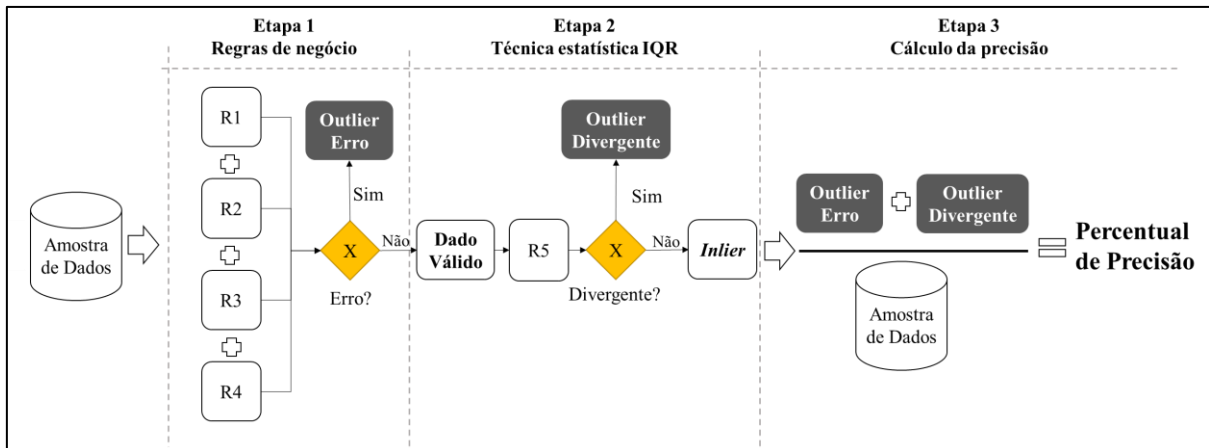


Figura 4. Aplicação do método proposto para identificação de *outliers* e cálculo da precisão

Fonte: Elaborado pelos autores.

Aguinis et al. (2013) sugerem que na avaliação de qualidade dos dados sejam utilizadas técnicas que intercalem abordagens quantitativas precedidas de ferramentas visuais; desta forma, com intuito de realizar a exame visual dos dados, neste trabalho foi utilizado o gráfico de *box plot*. Este tipo de gráfico foi criado por Tukey (1977) utilizando a técnica IQR; nele os grupos de dados são representados pelos seus quartis; as linhas que se estendem a partir das caixas apresentam valores que vão do mínimo ao máximo, podendo indicar a presença de *outliers* quando os limites das linhas estiverem muito distantes das caixas. Também foram utilizados gráficos de dispersão; criado por Francis Galton no início do século 20, este tipo de representação permite exibir graficamente valores de um conjunto de dados de duas (ou mais) variáveis independentes (Galton, 1886).

Com intuito de validar a eficácia do método proposto, além das análises visuais, foi calculado, por meio da regressão linear simples, o coeficiente de determinação (R^2) utilizado

para demonstrar o ajuste das variáveis antes e depois da extração dos dados com erros e divergentes; valores de R^2 próximos de 1 demonstram que os dados representam melhor as variáveis que os compõe e conseqüentemente o processo no qual foram gerados.

4. RESULTADOS

Os dados utilizados neste trabalho compreendem uma amostra do período de 01 de janeiro a 31 de dezembro de 2019, totalizando 67.983 registros coletados no processo. Considerando que as médias de consumo de combustível variam em função das classes de veículos, os dados foram divididos e plotados de acordo com as quatro classes: Comum, Padron, SemiPadron e Articulados. ABNT (2009).

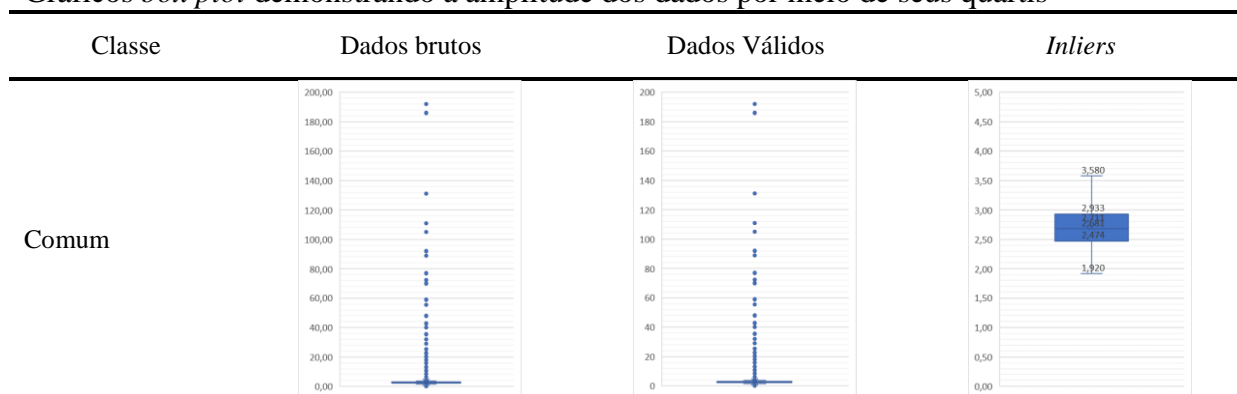
Os gráficos *box plot* permitiram avaliar a amplitude e distribuição dos limites entre os quartis em três cenários distintos: dados brutos, dados válidos e *inliers*, conforme apresentado na Tabela 3; nesta tabela, tanto na coluna de dados brutos quanto dados válidos, identifica-se a existência de valores que se afastam demasiadamente das bordas do 1º e principalmente do 3º quartil, e a obliquidade (falta de simetria da distribuição), desta forma, nesta primeira análise visual já é possível identificar uma amostra com apontamentos de dados imprecisos, pois estão muito além ou aquém da média geral de consumo mínimo e máximo das respectivas classes de veículos. Após a aplicação do conjunto das cinco regras os gráficos *box plot*, apresentados na coluna *inliers*, demonstram limites e divisões de quartis muito mais próximos da centralidade e da realidade constatada nas empresas para cada classe de veículo.

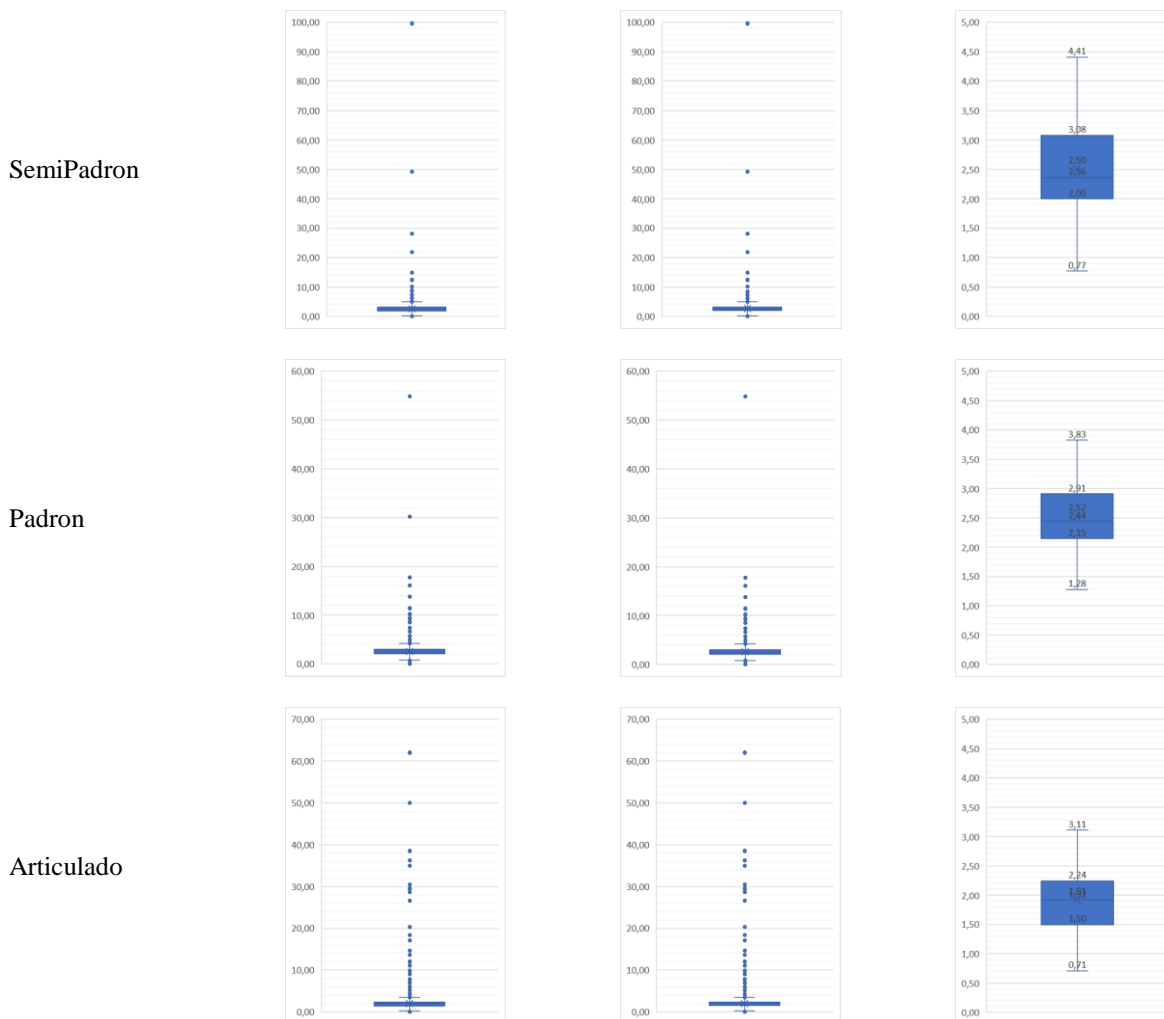
Outro ponto que pode ser destacado como negativo na precisão da qualidade dos dados ao avaliar os gráficos *box plot*, é a variação de amplitude, mesmo no conjunto de dados

inliers; por exemplo, na classe de veículos Comum, que possui a menor amplitude do IQR, o limite mínimo do *whisker* (1,92) está 28,38% abaixo da mediana (2,68) e o limite máximo (3,58) está 33,58% acima, e esta grande variação se repete nas demais classes; isto indica que, mesmo após aplicadas todas as regras de eliminação de erros e divergências, os dados ainda possuem um elevado grau de dispersão. Uma nova aplicação da regra R5 sobre o conjunto de dados *inliers* reduziria estes limites, todavia isto reduz ainda mais os apontamentos de dados na amostra avaliada, pois de acordo com Chandola et al., (2009) a região entre $Q1 - 1,5 IQR$ e $Q3 + 1,5 IQR$ já contém 99,3% das observações.

As Tabelas 3 e 4 apresentam os dados agrupados em colunas que indicam os resultados de acordo com cada etapa de aplicação do método proposto neste estudo. A coluna de dados brutos é composta pelas informações originalmente coletadas, sem nenhum tratamento; os dados válidos são resultado da aplicação das regras de negócio, R1, R2, R3 e R4 para eliminação dos dados com erros e na coluna *inliers* os gráficos apresentam dados após a aplicação das cinco regras.

Tabela 3
Gráficos *box plot* demonstrando a amplitude dos dados por meio de seus quartis





Fonte: Dados da pesquisa.

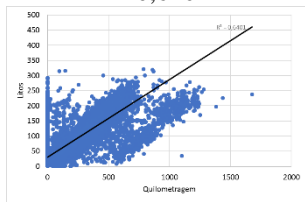
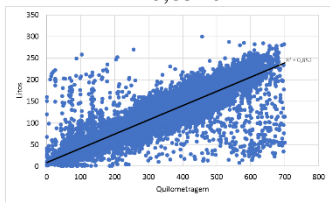
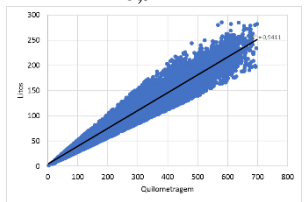
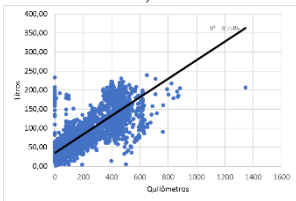
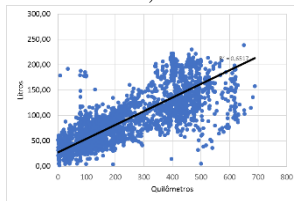
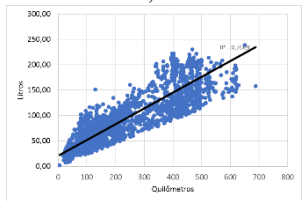
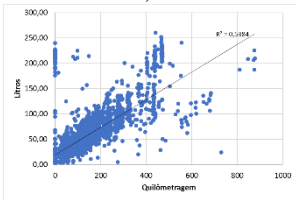
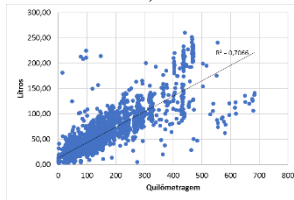
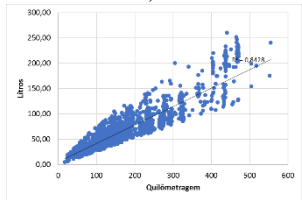
Os gráficos apresentados na Tabela 4 demonstram a dispersão dos pontos de dados plotados por meio de coordenadas cartesianas (eixo X, dados da quilometragem percorrida; eixo Y, dados dos litros de combustível consumido); junto de cada gráfico é apresentado também o respectivo coeficiente de determinação R^2 .

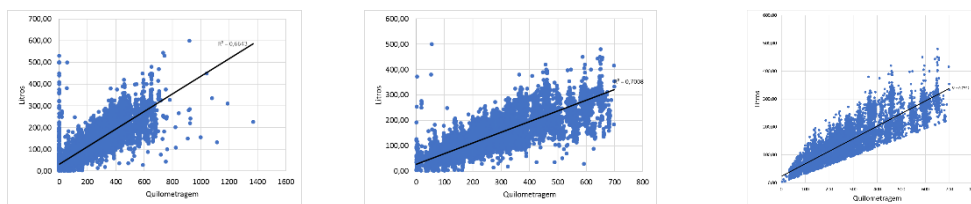
Durante a análise visual é possível constatar que nos gráficos da classe Comum verifica-se que os dados brutos possuem muitos registros com dispersões significativas, principalmente no que diz respeito a quilometragem máxima diária, caracterizando várias infrações aos limites propostos pelas regras de negócio. Ainda que o gráfico de dispersão desta categoria demonstre na coluna *inliers* pontos muito mais concentrados ao redor da reta,

e conseqüentemente R^2 muito próximo de 1, isto é ocasionado pelo fato desta ser a amostra onde foram eliminados o maior número de *outliers* divergentes; isto melhorou de maneira significativa o coeficiente de determinação, porém, reduziu o percentual de precisão.

Este comportamento relativo a maior amostra de dados (Comum) possuir o melhor coeficiente de determinação, porém, o menor resultado na precisão dos dados não é regular, afinal, a segunda maior amostra (Articulados) não possui o segundo melhor coeficiente de determinação, mas possui a maior precisão dos dados; isto demonstra que a quantidade de apontamentos de dados que se afastam da reta no cálculo da regressão exercem maior influência na precisão, do que apontamentos que excedem de maneira expressiva os limites do IQR.

Tabela 4
Gráficos de dispersão dos dados e coeficientes de determinação (R^2)

Classe	Dados Brutos	Dados Válidos	<i>Inliers</i>
Comum	<p>R^2 0,6401</p> 	<p>R^2 0,8520</p> 	<p>R^2 0,9411</p> 
SemiPadron	<p>R^2 0,5586</p> 	<p>R^2 0,6517</p> 	<p>R^2 0,7669</p> 
Padron	<p>R^2 0,5384</p> 	<p>R^2 0,7066</p> 	<p>R^2 0,8428</p> 
Articulado	<p>R^2 0,6643</p>	<p>R^2 0,7008</p>	<p>R^2 0,7722</p>



Fonte: Dados da pesquisa.

A contagem dos pontos de dados com erros (identificados pelas regras R1 a R4) e dos *outliers* divergentes (identificados pela regra R5) possibilitou calcular a representatividade percentual destes dados sobre o total de registros da amostra avaliada; o percentual de precisão geral obtido foi de 87,36%; este resultado é relativamente baixo quando comparado a classe Articulado, por exemplo, a qual apresentou a maior precisão (91,70%), mas ainda assim, está abaixo de níveis de precisão comumente aceitos em testes de amostras, por exemplo, os quais espera-se um valor mínimo de 95%. Uma das causas deste resultado relativamente baixo, pode ser identificada ao avaliar o percentual geral de dados com erros (2,90%) o qual está muito acima do percentual da classe Articulado (1,16%); isto se deve ao fato dos dados com erros se centrarem principalmente na classe Comum, a qual tem a maior representatividade sobre a amostra. Outro destaque diz respeito ao percentual de dados com erros das classes com maior número de registros de dados na amostra; na Comum, o percentual está mais de três vezes acima do percentual de erro da classe Articulado, a qual detém o segundo maior tamanho de amostra. Isto demonstra que a acuracidade do processo de coleta e tratamento de dados não está relacionada apenas ao volume de veículos abastecidos e dados coletados. A Tabela 5 apresenta o resumo dos resultados quantitativos e percentuais obtidos.

Tabela 5

Resumo dos resultados da aplicação do método nos dados coletados

Classe	Registros	Erros	Válidos	% Erro	<i>Outlier</i>	<i>Inlier</i>	% <i>Outlier</i>	Precisão Final
Articulado	17.504	203	17.301	1,16%	1.249	16.052	7,22%	91,70%
Padron	4.318	71	4.247	1,64%	343	3.904	8,08%	90,41%
SemiPadron	3.768	100	3.668	2,65%	370	3.298	10,09%	87,53%
Comum	42.393	1.599	40.794	3,77%	4.657	36.137	11,41%	85,25%
Total	67.983	1.973	66.010	2,90%	6.619	59.391	10,02%	87,36%

Fonte: Dados da pesquisa.

Com intuito de avaliar o grau de impacto sobre o coeficiente de determinação quando da eliminação de dados com erros e *outliers* da amostragem, a Tabela 6 exhibe os resultados desta inspeção; na tabela constata-se que os dados brutos não possuem um coeficiente de determinação suficientemente representativo, o qual comumente deve estar acima de 0,7. Porém, a partir da eliminação de dados com erros o coeficiente de determinação ultrapassa o limite mínimo de 0,7 em três das quatro classes, destacando a melhora expressiva (33,10%) da classe Comum. A eliminação de *outliers* resultada em uma nova melhora na adequação dos valores em relação às variáveis, o que se percebe com o aumento expressivo (14,1%) do coeficiente médio de determinação. Em todas as classes de veículos os dados resultantes da aplicação das regras apresentam coeficiente de determinação acima de 0,7, destacando novamente como positivo o resultado da classe Comum (0,9411), no qual os pontos de dados resultantes estão fortemente ajustados à reta.

Tabela 6
Variação do coeficiente de determinação (R^2)

Classe	R^2 Dados Brutos	R^2 Dados Válidos	Variação	R^2 <i>Inliers</i>	Variação
Comum	0,6401	0,8520	33,10%	0,9411	10,5%
SemiPadron	0,5586	0,6517	16,67%	0,7669	17,7%
Padron	0,5384	0,7066	31,24%	0,8428	19,3%
Articulado	0,6643	0,7008	5,49%	0,7722	10,2%
Média	0,6004	0,7278	21,23%	0,8308	14,1%

Fonte: Dados da pesquisa

5. DISCUSSÃO DOS RESULTADOS

Mesmo de forma empírica, a avaliação da QD compõe uma das habilidades dos tomadores de decisão (Earley & Henderson, 2017); compreender até que ponto podem confiar nos valores de dados subjacentes para apoiar suas decisões é melhor assegurado quando a incerteza destes dados é verificada por meio de métricas de qualidade bem fundamentadas. Heinrich et al., 2018); neste trabalho foi possível obter esta medida por meio do cálculo da dimensão da precisão.

A combinação na utilização de regras híbridas, ou seja, regras de negócios fundamentais do processo (Barbieri & Farinelli, 2013) e técnica estatística (IQR), demonstrou que o método utilizado é capaz de identificar e segregar dados de baixa qualidade em diferentes cenários; mesmo dados que cumprem regras de negócios estabelecidas pela organização, mas que se desviam de forma excessiva de resultados considerados factuais para o processo mapeado e que, segundo Zoet et al., (2011) caracterizam quebras de conformidade.

No contexto avaliado não foi identificada correlação linear entre o tamanho da amostra e o resultado percentual de precisão dos dados. Considerando que uma frota de veículos e as características do TPC pode ser bastante heterogênea, (Kuah; Perl, 1988) pois depende da população, da demanda e da quantidade de empresas que atuam em uma região, este comportamento é positivo, pois demonstra que o método não favorece ou prejudica a avaliação da precisão de amostras em função de seus tamanhos, podendo ser adotado em qualquer contexto de negócio ou processos distintos.

A eliminação dos pontos de dados com erro (R1 a R4) melhorou, em média, 21,63% o coeficiente de determinação, demonstrando que os dados válidos apresentavam melhor ajuste em relação às variáveis independentes; porém não eliminou pontos extremos dos intervalos interquartis. A aplicação da técnica IQR gerou, em média, mais 14,40% de melhora do coeficiente de determinação e eliminando os pontos extremos, demonstrando também a eficácia do método proposto na identificação de registros que interferem na precisão do processo.

Durante a execução desta atividade constatou-se que as regras de negócio presentes no sistema ERP, com intuito de impedir a gravação de dados incorretos, não foram eficazes, pois vários apontamentos de dados ultrapassaram os parâmetros; isto denota a importância em garantir que dados mestres no sistema ERP sejam controlados, mantidos e inspecionados periodicamente, conforme preconiza o DAMA-BOK®, assim como, a adoção de uma metodologia permanente de medição da qualidade dos dados.

Outras implicações quanto a geração de dados incorretos neste processo, são a interferência no controle do estoque de combustível, que serve de subsídios para as decisões de compras; e, no caso das distâncias percorridas, afetam a programação de manutenção dos veículos.; assim, as consequências de decisões erradas se tornam cada vez mais caras (Forbes

Insights, 2017). Uma vez que o processo de TPC é contínuo, estas implicações podem ocorrer constantemente, então a inspeção dos dados em intervalos diários é recomendada pelo DAMA-BOK®, e neste caso reforçada pela presença de dados extremos na amostra, como quilometragem ou consumo muito acima daqueles previstos nas regras de negócios.

6. CONSIDERAÇÕES FINAIS

O crescimento expressivo no volume e a necessidade de estar em conformidade com as regras dos processos empresariais, tem destacado a importância de trabalhar com dados com qualidade. Neste artigo foi apresentado um método de medição de qualidade dos dados, utilizando regras de negócio e técnica estatística, avaliando a dimensão da precisão; a obtenção do percentual de precisão permitiu identificar com qual exatidão o conjunto de dados representou o processo no qual foi gerado.

O modelo foi testado utilizando dados reais do processo de consumo e abastecimento de combustível de empresas de transporte público coletivo da região de Curitiba-PR, com a aplicação de regras do negócio para identificação de erros e utilização da técnica de cálculo dos intervalos interquartis para identificação de *outliers*.

A partir do mapeamento do processo foram coletados e verificados 67.983 registros, dos quais foram identificados 1.973 *outliers* de erros e 6.619 *outliers* divergentes, resultando em uma precisão de 87,36% dos dados da amostra. Estes números demonstram que o método apontou, que os dados coletados não representam o processo de consumo e abastecimento de combustível com exatidão; contribuiu inclusive para evidenciar, de forma bastante clara, a presença de apontamentos de dados discrepantes com as regras de negócio instituídas, levando os gestores a utilizar dados de qualidade reduzida em seus processos de decisão.

Como contribuição geral este trabalho apresenta uma metodologia de avaliação de qualidade de dados, com base na dimensão da precisão, por meio da localização de informações que não expressam fielmente o processo real, além de validar a eficácia do método com a aplicação do cálculo do coeficiente de determinação em diferentes cenários; ao meio empresarial este método possibilita a melhoria da acuracidade das decisões, a partir da

consciência da precisão dos dados coletados no processo, assim como, a motivação para a criação de um programa formal de gerenciamento de qualidade dos dados.

Este trabalho se limitou a avaliar a precisão dos dados de um processo, em uma amostra de empresa que utilizam o mesmo sistema ERP e o mesmo processo de geração e tratamento de dados; desta forma, a continuidade desta pesquisa pode aplicar este método em dados obtidos por meio de outras ferramentas de coleta e tratamento de dados, como a Internet das Coisas (IoT), assim como, ser aplicada em outros processos de negócios; com isso, demonstrar se o grau de precisão dos dados é influenciado pelo processo de coleta e tratamento.

REFERÊNCIAS

- ABNT. Associação Brasileira de Normas Técnicas. NBR 15570 “Transporte — Especificações técnicas para fabricação de veículos de características urbanas para transporte coletivo de passageiros”. 2009.
- Aguinis, H.; Gottfredson, R. K. & Joo, H. (2013) Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods*, v. 16, n. 2, p. 270–301.
- Anderson, D. R.; Sweeney, D. J.; Williams, T. A.; Camm, J. D.; Cochran, J. J. (2016). *Statistics for business & economics*. Nelson Education.
- Barbieri, C.; Farinelli, F. (2013). Uma visão sintética e comentada do Data Management Body of Knowledge (DMBOK).
- Bizer, C.; Cyganiak, R. (2009). Quality-driven information filtering using the WIQA policy framework. *Web Semantics*, v. 7, n. 1, p. 1–10.
- Brynjolfsson, E.; Hitt, L. & Kim, H. (2011). Strength in numbers: How does data-driven decision-making affect firm performance? *International Conference on Information Systems 2011, ICIS 2011*, v. 1, p. 541–558, 2011.
- Data Management Association (dama) uk working group on “data quality dimensions”. (2013). *The six primary dimensions for data quality assessment: defining data quality dimensions*. UK.
- Earley, S.; henderson, D. (2017). *DAMA-DMBOK: data management body of knowledge*. 2^o ed. New Jersey: Bradley Beach.
- Even, A.; Shankaranarayanan, G. (2007). Utility-Driven Assessment of Data Quality. *Data Base for Advances in Information Systems*, v. 38, n. 2, p. 75–93.
- Forbes Insights. (2017). *The Data Differentiator. How Improving Data Quality Improves Business*. Forbes Media, New York.
- Galton, F. (1986). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, v. 15, p. 246–263
- Hawkins, D. M. (1980). *Identification of Outliers*. London ; New York: Chapman and Hall, 1980.

- Heinrich, B.; Hristova, D.; Klier, M.; Schiller, A.; Szubartowicz, M. (2018). Requirements for data quality metrics. *Journal of Data and Information Quality (JDIQ)*, v. 9, n. 2, p. 1–32. ACM New York, NY, USA.
- Kuah, G. K.; Perl, J. (1988). Optimization of feeder bus routes and bus-stop spacing. *Journal of Transportation Engineering*, v. 114, n. 3, p. 341–354. American Society of Civil Engineers.
- NTU. Associação Nacional das Empresas de Transportes Urbanos. Transporte público como direito social. E agora? (2016) Brasília: NTU.
- NTU. Associação Nacional das Empresas de Transportes Urbanos. Anuário NTU 2017-2018. Brasília: NTU.
- Prasad, N. R.; Almanza-Garcia, S.; LU, T. T. (2009). Anomaly detection. *Computers, Materials and Continua*, v. 14, n. 1, p. 1–22.
- Price, R.; Shanks, G. A. (2005) Semiotic Information Quality Framework: Development and Comparative Analysis. *Journal of Information Technology*, v. 20, n. 2, p. 88–102, SAGE Publications Ltd. Disponível em: <<https://doi.org/10.1057/palgrave.jit.2000038>>. .
- Shankaranarayan, G.; Ziad, M.; Wang, R. Y. (2003). Managing data quality in dynamic decision environments: An information product approach. *Journal of Database Management*, v. 14, n. 4, p. 14–32.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Princeton: Addison-wesley Longman, Inc,
- URBS. Urbanização de Curitiba S/A. (2021). Categorias de Linhas. Disponível em: <<https://www.urbs.curitiba.pr.gov.br/transporte/rede-integrada-de-transporte/24>>. Acesso em: 13/1/2021.
- URBS. Urbanização de Curitiba S/A. (2020) Custos da Rede Integrada de Transporte. Disponível em: <<http://www.urbs.curitiba.pr.gov.br/transporte/tarifas-custos>>. Acesso em: 1/3/2020.
- Vinutha, H. P.; Poornima, B.; Sagar, B. M. (2018). Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset. In: Satapathy S., Tavares J., Bhateja V., Mohanty J. *Information and Decision Sciences. Advances in Intelligent Systems and Computing*. p.511–518. Singapore: Springer.
- Wang, R.; Pierce, E. M.; Madnick, S.; Fisher, C. (2005). *Information Quality*. 1^o ed. New York: Routledge.
- Wang, R. Y.; Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, v. 12, n. 4, p. 5–33. Taylor & Francis.
- Zoet, M.; Versendaal, J.; Ravesteyn, P.; Welke, R. (2011). Alignment of Business Process Management and Business Rules. *ECIS 2011 Proceedings*, v. 34. Disponível em: <<https://aisel.aisnet.org/ecis2011/34/>>.