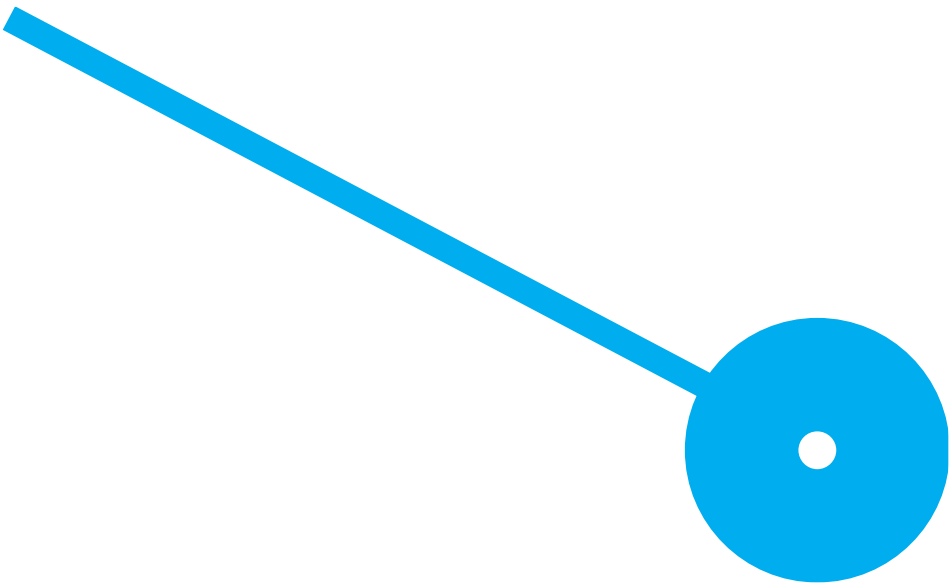
ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

MESTRADO
ENGENHARIA INFORMÁTICA

Análise e Monitorização da Qualidade de Dados no contexto da Indústria 4.0

Teresa Maria Oliveira Peixoto

07/2025



ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

M MESTRADO
ENGENHARIA INFORMÁTICA

Análise e Monitorização da Qualidade de Dados no contexto da Indústria 4.0

Teresa Maria Oliveira Peixoto
8190334

Orientadores

Prof. Doutor Bruno Moisés Teixeira de Oliveira
Prof. Doutor Óscar António Maia de Oliveira

Dissertação apresentada para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática pela Escola Superior de Tecnologia e Gestão do Instituto Politécnico do Porto.

07/2025

Declaração de integridade

Eu, Teresa Maria Oliveira Peixoto, estudante nº 8190334, do Mestrado em Engenharia Informática da Escola Superior de Tecnologia e Gestão do Instituto Politécnico do Porto, declaro que não fiz plágio nem auto-plágio, pelo que o trabalho intitulado "Análise e Monitorização da Qualidade de Dados no contexto da Indústria 4.0" é original e da minha autoria, não tendo sido usado previamente para qualquer outro fim. Mais declaro que todas as fontes usadas estão citadas, no texto e na bibliografia final, segundo as regras de referenciação adotadas na instituição.

Agradecimentos

Esta dissertação marca o fim de uma etapa desafiante e enriquecedora, possível apenas com o apoio e presença de várias pessoas e entidades, a quem deixo o meu sincero agradecimento.

Aos meus pais, um agradecimento especial e incondicional, sem o vosso apoio, amor e sacrifício, nada disto teria sido possível. Obrigada por me apoiarem, por darem os melhores conselhos e por me ajudarem a crescer. Estarei para sempre grata!

Às minhas irmãs, por me desafiarem de forma única, mas também por me ensinarem o verdadeiro significado de união, apoio e amizade.

À minha tia e ao meu tio, por estarem sempre ao meu lado com palavras de incentivo, carinho e uma presença constante que tanto me confortou. Obrigada por acreditarem em mim e por transformarem até as pequenas vitórias em grandes memórias de alegria.

Ao meu namorado, agradeço por nunca ter deixado de estar ao meu lado, mesmo nos dias em que me sentia perdida. Obrigada por me abraçares com paciência quando tudo parecia demasiado, por me lembrares de parar, respirar e seguir com calma, e por trazeres sempre a tua luz nos momentos em que mais precisei. O teu apoio não só me ajudou a resistir, mas também me deu coragem, clareza e a certeza profunda de que tudo era possível.

À Salsicheira, a minha cadela, que mesmo a exigir atenção a toda a hora, nunca deixou de estar por perto. A sua presença trouxe companhia e aquele carinho silencioso que, sem saber, me ajudou a manter o equilíbrio nos dias mais desafiantes.

Aos meus amigos, obrigada por estarem sempre presentes, mesmo longe e nos momentos em que o tempo e o cansaço apertavam. Obrigada pelas conversas genuínas, pelas gargalhadas espontâneas e por me ajudarem a desligar quando precisava. O vosso apoio, mesmo em pequenos gestos ou mensagens fora de horas, fez toda a diferença.

Aos meus orientadores, professor Dr. Bruno Oliveira e professor Dr. Óscar Oliveira, agradeço profundamente pela orientação, pela disponibilidade constante e pelo incentivo à superação. O vosso apoio foi determinante para a qualidade e solidez deste trabalho.

A todos os docentes da ESTG, o meu agradecimento pelo ambiente de excelência e, sobretudo, pelos conhecimentos e experiências que me transmitiram ao longo destes anos.

Ao CIICESI, pela bolsa de investigação, e à FCT, pelo financiamento atribuído no âmbito do projeto PRODUTECH R3, integrado no Plano de Recuperação e Resiliência, com apoio da União Europeia através do Next Generation EU, essencial para o sucesso desta investigação e para a consolidação dos resultados alcançados.

Abstract

Industry 4.0 represents a paradigm shift in production systems, marked by the integration of advanced digital technologies such as the Internet of Things, Cyber-Physical Systems, and Artificial Intelligence. These technologies enable the massive, real-time collection of data, allowing for monitoring, automation, and optimization of industrial processes with unprecedented levels of precision and efficiency. However, the value of generated data inherently depends on its quality. Incomplete, inaccurate, inconsistent, or outdated data can compromise not only the reliability of analytical insights but also the safety and effectiveness of the decisions based on them.

In this context, data quality emerges as a central and strategic concern. Its assessment and continuous monitoring are essential to ensure that the collected data is fit for purpose, enabling early fault detection, inefficiency reduction, and system adaptation to actual operating conditions. This dissertation focuses on the analysis of the main data quality dimensions and metrics applied to industrial environments, exploring their use in real scenarios where data is continuously produced by sensors distributed across the value chain.

Quantitative approaches to measure quality dimensions such as accuracy, completeness, consistency, and timeliness are discussed, along with strategies to integrate these evaluations into continuous data streams. The study demonstrates that systematically and automatically incorporating data quality control mechanisms has a significant impact on improving the resilience, trustworthiness, and operational intelligence of modern industrial systems.

Keywords: Data Quality, Industry 4.0, Quality Metrics, Data Profiling, Continuous Monitoring

Resumo

A Indústria 4.0 representa uma mudança de paradigma nos sistemas de produção, caracterizada pela integração de tecnologias digitais avançadas, como a Internet das Coisas, os Sistemas Ciberfísicos e a Inteligência Artificial. Estas tecnologias possibilitam a recolha massiva de dados em tempo real, permitindo monitorizar, automatizar e otimizar processos industriais com um nível de precisão e eficiência sem precedentes. No entanto, o valor dos dados gerados depende intrinsecamente da sua qualidade. Dados incompletos, imprecisos, inconsistentes ou desatualizados podem comprometer não apenas a fiabilidade das análises, mas também a segurança e eficácia das decisões tomadas com base nesses dados.

Neste contexto, a qualidade dos dados assume um papel central e estratégico. A sua avaliação e monitorização tornam-se essenciais para garantir que os dados recolhidos sejam adequados ao propósito a que se destinam, permitindo a deteção precoce de falhas, a redução de ineficiências e a adaptação dos sistemas às condições reais de operação. Esta dissertação centra-se na análise das principais dimensões e métricas de qualidade aplicadas ao contexto industrial, explorando a sua aplicação em ambientes reais onde os dados são produzidos continuamente por sensores distribuídos ao longo da cadeia de valor.

São discutidas abordagens quantitativas para medir dimensões da qualidade como acurácia, completude, consistência e atualidade, bem como estratégias para integrar estas avaliações em fluxos contínuos de dados. A investigação demonstra que a incorporação de mecanismos de controlo da qualidade dos dados, de forma sistemática e automatizada, tem um impacto significativo na melhoria da resiliência, confiança e inteligência operacional dos sistemas industriais modernos.

Palavras-chaves: Qualidade dos Dados, Indústria 4.0, Métricas de Qualidade, *Data Profiling*, Monitorização Contínua

Conteúdo

Lista de Figuras

Lista de Tabelas

Acrónimos

DQD Dimensões da Qualidade dos Dados.

I4.0 Indústria 4.0.

IA Inteligência Artificial.

IdC Internet das Coisas.

LSTM *Long Short-Term Memory*.

LWQS *Longitudinal Weighted Quality Score*.

MAD Desvio Absoluto da Mediana.

QSD *Quality Score Delta*.

SCF Sistemas Ciberfísicos.

STL *Seasonal-Trend decomposition using Loess*.

WQS *Weighted Quality Score*.

Capítulo 1

Introdução

A transformação digital tem vindo a redefinir o panorama industrial nas últimas décadas, promovendo uma mudança significativa na forma como os processos produtivos são planeados, executados e otimizados. No centro desta revolução está a digitalização e a interconectividade dos sistemas produtivos, onde os dados assumem um papel estratégico e central (6).

Este novo paradigma apoia-se em diversas tecnologias emergentes, como a Internet das Coisas (IdC), os Sistemas Ciberfísicos (SCF), a Inteligência Artificial (IA), entre outras (7). A IdC permite estender a Internet ao mundo físico, conectando sensores e dispositivos ao ambiente digital, enquanto os SCF asseguram uma integração coesa entre componentes físicos e computacionais. Em conjunto, estas tecnologias possibilitam a recolha e transmissão de dados em tempo real, de forma autónoma, distribuída e eficiente (8).

A conectividade contínua tornou-se tão relevante quanto a gestão de materiais e produtos na indústria moderna (6). Com sensores distribuídos ao longo da cadeia de valor, os ambientes industriais geram volumes massivos de dados, estruturados e não estruturados, a uma velocidade sem precedentes (9). Frequentemente organizados como séries temporais, estes dados permitem uma monitorização constante e uma resposta rápida aos eventos no chão de fábrica (10). A baixa latência e a fiabilidade dos dados são, por isso, requisitos essenciais para a operação eficiente de sistemas industriais de Indústria 4.0 (I4.0) em tempo real, onde a gestão eficaz de grandes volumes de dados com continuidade e qualidade é fundamental para gerar valor e conhecimentos práticos (8).

Com o avanço da automação dos processos produtivos, a intervenção humana tem vindo a reduzir-se progressivamente, exigindo que os sistemas funcionem de forma mais autónoma e baseiem as suas decisões na análise dos dados disponíveis (8). Para isso, os sistemas devem ser capazes de processar grandes volumes de dados com elevada precisão e continuidade (8). Assim, as tecnologias de transmissão, análise e monitorização de dados tornaram-se pilares da capacidade de adaptação e resposta das organizações industriais aos seus contextos operacionais. Contudo, a eficácia dessas decisões está diretamente dependente da qualidade dos dados utilizados. À medida que aumentam a conectividade e a complexidade dos sistemas, aumentam também os desafios associados à gestão e fiabilidade dos dados (9). Não é possível confiar exclusivamente nos sensores para garantir dados de qualidade, dado que muitos são suscetíveis a erros, como falhas, interferências eletromagnéticas, perda de pacotes ou discrepâncias semânticas entre fontes (11). Estes

problemas podem comprometer a integridade da informação recolhida (11).

Apesar do seu enorme potencial para otimizar operações, reduzir custos e antecipar falhas, os dados só se tornam realmente úteis se garantirmos a sua qualidade, integridade e fiabilidade. Dados comprometidos, como valores em falta, ruído, duplicação de registos, inconsistências semânticas ou estruturas divergentes entre fontes são bastante recorrentes (1). Sem uma arquitetura escalável capaz de extrair, avaliar e melhorar a qualidade dos dados, as empresas tendem a acumular grandes volumes de dados que se tornam difíceis de explorar e converter em valor real. Isto resulta no que é conhecido como "dados obscuros", o que compromete análises e decisões, originando custos operacionais e riscos significativos. Segundo Corallo et al. (12), dados obscuros são dados não catalogados ou mal estruturados que são gerados, recolhidos e armazenados durante as operações, mas que não são analisados devido à falta de ferramentas analíticas adequadas e acabam por ser ignorados pelas organizações (9).

Nesse sentido, a qualidade dos dados emerge como um dos principais desafios e, simultaneamente, como uma das áreas mais críticas a abordar no contexto da I4.0. A diversidade de fontes e formatos de dados constitui um desafio significativo para garantir a qualidade dos dados em tempo real, tornando essencial a utilização de métodos avançados de monitorização (1). Um dos principais obstáculos está na heterogeneidade das fontes de dados, sobretudo na integração entre sensores industriais e sistemas externos (13). A combinação de dados de múltiplos sensores, com diferentes características técnicas, escalas e formatos, exige técnicas robustas para assegurar que os dados integrados sejam fiáveis e úteis (14).

A avaliação da qualidade dos dados torna-se, assim, uma necessidade estratégica para todos os tipos de organizações, permitindo o progresso contínuo e minimizando, ou mesmo eliminando, problemas relacionados com a integridade dos dados (15), onde é necessário compreender se os dados são apropriados ao seu proposito (9). Esta avaliação é guiada por um conjunto de dimensões reconhecidas na literatura, sendo as mais recorrentes: acurácia, completude, consistência e atualidade (16). Estas dimensões representam perspetivas complementares sobre o que significa qualidade num dado contexto, e estão associadas a métricas específicas que as quantificam de forma objetiva. O uso de métricas é essencial para distinguir entre dados de alta e baixa qualidade, com base na sua utilidade prática (9). Contudo, apesar da grande quantidade de propostas, ainda não existe um quadro comum para a avaliação padronizada da qualidade dos dados industriais (17). A seleção e aplicação de métricas adequadas para cada dimensão é, por isso, um fator essencial para garantir uma avaliação rigorosa e operacionalmente relevante.

Ainda assim, é necessário a adoção de técnicas, como o *data profiling*, que permite identificar vários problemas nos dados de forma contínua e com mínima intervenção humana (1). Através da definição de *data profiling*, os engenheiros e analistas de dados obtêm uma visibilidade detalhada de vários atributos de dados, tais como distribuições, tipos de dados, e relações entre tabelas ou bases de dados (18). Esta visão ajuda a estabelecer métricas da qualidade, permitindo que as equipas definam padrões de qualidade realistas e significativos (3).

Deste modo, a presente dissertação procura responder à seguinte questão de investigação:

> *De que forma se pode analisar, avaliar e monitorizar, em tempo real, a qualidade dos dados em ambientes industriais?*

Para melhor enquadrar esta problemática, a questão principal pode ser desdobrada nas

seguintes subquestões:

- Que tipos de problemas de qualidade dos dados podem ser detetados em contextos industriais e de que forma estes impactam os processos analíticos?
- Como podem ser aplicadas, e quais, as métricas e técnicas de *data profiling* mais adequadas para identificar dados de baixa qualidade?
- De que forma é possível monitorizar, em tempo real, os dados e a sua qualidade, tendo em conta os requisitos da Indústria 4.0 e dos ambientes Big Data?

1.1 Objetivos

O objetivo central desta dissertação é propor, implementar e validar uma abordagem prática, escalável e automatizada para a monitorização contínua da qualidade dos dados em ambientes industriais, recorrendo a dimensões e métricas amplamente reconhecidas na literatura. Pretende-se que esta abordagem contribua para o desenvolvimento de processos mais robustos, fiáveis e resilientes, alinhados com os requisitos operacionais da I4.0.

Mais concretamente, este trabalho visa:

- Identificar e sistematizar as principais dimensões e métricas utilizadas na avaliação da qualidade de dados em contextos industriais;
- Aplicar essas métricas a dados de sensores industriais em funcionamento contínuo;
- Aplicar um índice de qualidade de dados que permita uma avaliação agregada e temporal;
- Implementar sistema de monitorização contínuo, desde a ingestão dos dados até à visualização, que suporte decisões operacionais baseadas em dados de elevada qualidade.

1.2 Metodologia

Este trabalho adota uma metodologia baseada em três fases sequenciais e iterativas: revisão do domínio, análise através de casos de estudo e desenvolvimento da solução. Estas fases são complementares entre si e estruturam o trabalho desde a compreensão inicial do problema até à construção e validação de uma solução prática.

O projeto teve início com o estudo aprofundado do domínio, procurando compreender o funcionamento dos sistemas industriais, os requisitos operacionais, os principais desafios e as limitações mais comuns associadas à monitorização de dados neste tipo de ambiente. Esta fase incluiu também uma revisão sistemática da literatura, com o objetivo de identificar os principais conceitos, como dimensões e métricas associadas à qualidade dos dados no contexto da I4.0. O conhecimento adquirido serviu de base para as decisões técnicas subsequentes.

Com base nos conhecimentos obtidos, foram analisados diferentes conjuntos de dados provenientes de sensores industriais. Aplicaram-se diversas métricas de qualidade, abordagens baseadas em janelas temporais e filtros estatísticos, com o objetivo de validar conceitos e avaliar a eficácia e estabilidade das soluções propostas. Sempre que necessário,

recorreu-se à fase anterior, para clarificar aspetos do domínio ou reavaliar decisões metodológicas.

Na última fase, foi concebido e implementado um *pipeline* de qualidade dos dados, capaz de realizar a ingestão, validação, análise e monitorização de dados em tempo real. A arquitetura incluiu a integração de ferramentas e o desenvolvimento de componentes para o cálculo de métricas e indicadores agregados de qualidade. Esta implementação permitiu operacionalizar os conceitos teóricos e obter uma infraestrutura funcional e reutilizável.

1.3 Estrutura da Tese

A dissertação está estruturada em cinco capítulos principais:

- **Capítulo 1 – Introdução:** Contextualiza o problema, apresenta os desafios e define os objetivos da investigação;
- **Capítulo 2 – Estado da Arte:** Analisa os principais conceitos relacionados com a qualidade de dados, as dimensões e métricas mais relevantes, técnicas de *Data Profiling*, bem como sistemas e *pipelines* de monitorização utilizados na I4.0;
- **Capítulo 3 – Casos de Estudo:** Descreve e analisa os três diferentes casos de estudo realizados, com base em dados reais;
- **Capítulo 4 – Discussão:** Apresenta uma análise crítica e comparativa dos resultados obtidos, discutindo vantagens, limitações e contributos das abordagens adotadas;
- **Capítulo 5 – Conclusões:** Resume os principais resultados, evidencia as contribuições da dissertação e propõe direções para trabalhos futuros.

Os capítulos desta dissertação foram desenvolvidos com base em artigos científicos previamente publicados ou submetidos, no âmbito da investigação realizada com o apoio do **CIICESI**, através de bolsa atribuída no contexto do projeto **PRODUTECH R3**, integrado no Plano de Recuperação e Resiliência da República Portuguesa. A descrição detalhada desses contributos encontra-se no Capítulo 6 – Resultados Científicos.

Capítulo 2

Estado da arte

A qualidade dos dados é um fator essencial para o sucesso de qualquer iniciativa orientada por dados, especialmente em ambientes industriais. Embora a recolha e transmissão de dados em tempo real estejam amplamente disseminadas graças a tecnologias como IdC e SCF, a sua utilidade prática depende diretamente da integridade, consistência e adequação dos dados aos processos que os suportam (1).

Neste contexto, a literatura tem vindo a consolidar múltiplas abordagens à caracterização da qualidade dos dados, com ênfase crescente na monitorização contínua. A avaliação da qualidade de dados nos cenários da I4.0 não se limita à deteção de anomalias ou falhas pontuais, requer a definição e aplicação sistemática de dimensões e métricas capazes de quantificar, em tempo real, o grau de fiabilidade da informação recolhida (1). Assim, garantir a qualidade dos dados deixou de ser um requisito secundário para se tornar um elemento central da arquitetura digital.

Dimensões como acurácia, completude, consistência e atualidade são frequentemente referidas como pilares fundamentais para essa avaliação (16). Apesar da importância reconhecida destas dimensões, a sua operacionalização em contextos industriais levanta desafios substanciais. A heterogeneidade de fontes, a ausência de padrões comuns de representação e a elevada variabilidade dos dados implicam que as métricas aplicadas sejam sensíveis ao domínio, escaláveis e adaptáveis ao tempo (9). Além disso, uma parte significativa dos dados gerados permanece inexplorada, os chamados "dados obscuros", não por falta de relevância, mas por limitações tecnológicas ou metodológicas no seu aproveitamento (12).

Deste modo, este capítulo reúne, analisa e organiza criticamente o conhecimento atual sobre qualidade dos dados no contexto da I4.0. Apresenta-se uma revisão das principais abordagens à definição das Dimensões da Qualidade dos Dados (DQD), métodos de classificação, métricas propostas, desafios emergentes e técnicas de suporte como o *data profiling*. O objetivo é construir um referencial sólido que sirva de base para o desenvolvimento de soluções de monitorização da qualidade de dados em ambientes industriais reais.

A monitorização da qualidade dos dados tem sido um tema recorrente na literatura científica, especialmente no contexto de ambientes industriais inteligentes onde a fiabilidade dos dados influencia diretamente a eficácia da tomada de decisão. A seguir, são apresentadas e discutidas várias abordagens e contribuições relevantes que exploram os desafios e soluções aplicadas à qualidade dos dados no contexto da I4.0.

Rangineni et al. (19) realizaram uma revisão extensiva focada em como melhorar a qualidade dos dados para maximizar o desempenho da análise de dados em diferentes contextos organizacionais. Os autores destacam que as falhas de qualidade não apenas comprometem os resultados analíticos, mas também podem distorcer a perceção do negócio e dificultar a geração de valor a partir dos dados. A revisão agrupa os desafios e soluções em torno de três eixos principais: fontes de problemas de qualidade, técnicas de melhoria e estratégias de governação de dados. Para responder aos desafios identificados, o estudo sintetiza várias técnicas de melhoria da qualidade dos dados, destacando a limpeza de dados, o *data profiling*, a normalização e padronização e a análise preditiva, sendo salientada a importância de uma cultura organizacional orientada à qualidade como fator facilitador de melhorias sustentáveis. Os autores defendem que, para além de soluções técnicas, é necessário estabelecer políticas internas claras, papéis bem definidos e métricas operacionais de monitorização da qualidade. A literacia de dados nas equipas é igualmente apontada como essencial, dado que muitos problemas são intensificados por má interpretação ou uso inadequado da informação disponível. Por fim, o artigo sublinha que iniciativas de melhoria contínua da qualidade devem ser vistas como um investimento estratégico, e não como um custo.

Goknil et al. (9) conduziram uma revisão sistemática abrangente que incide sobre aplicações em ambientes SCF e IdC na I4.0, identificando desafios específicos na integração de dados de múltiplas fontes heterogéneas. O foco central da investigação está na forma como diferentes abordagens tecnológicas enfrentam os desafios de qualidade no contexto da I4.0, com ênfase particular em dimensões, métricas e técnicas automatizadas de melhoria de dados. Identificam-se como principais problemas de qualidade os valores em falta, outliers, ruído, inconsistências e imprecisões. Uma das conclusões mais críticas é que, apesar da existência de múltiplas métricas na literatura, estas raramente são implementadas em sistemas industriais reais. Além disso, os autores destacam a limitação de técnicas não baseadas em IA e defendem a necessidade de soluções orientadas por aprendizagem que permitam tanto reparação como limpeza de dados em tempo real, com suporte a diferentes camadas da arquitetura IoT. Um dos principais contributos do artigo é a classificação detalhada das técnicas de qualidade dos dados em três grandes categorias: monitorização dos dados, limpeza dos dados e reparação dos dados. Goknil et al. sistematizam 17 dimensões de qualidade de dados identificadas ao longo da literatura. Apesar desta riqueza conceitual, os autores observam que a maioria dos estudos aplica apenas um subconjunto reduzido destas dimensões, sem justificativa clara e raramente são utilizadas métricas formais para avaliar essas dimensões de forma objetiva.

Zhang et al. (20) abordam a gestão da qualidade de dados especificamente em ambientes IdC. O estudo apresenta um mapeamento das metodologias disponíveis, incluindo normas internacionais, e frameworks de avaliação baseadas em dimensões e reconhece que os dados provenientes de sensores são, por natureza, ruidosos, incompletos, heterogéneos e propensos a falhas. É dada particular atenção à integração dos dados e à aplicação de mecanismos de validação e agregação que permitam atenuar os efeitos de ruído e lacunas nos fluxos de dados. O artigo também identifica um conjunto robusto de dimensões de qualidade dos dados e destaca a necessidade de definir métricas específicas para cada dimensão, alertando que a maioria dos projetos IdC ignora a definição formal destas métricas, utilizando abordagens *ad hoc* ou reativas, o que limita a capacidade de identificar e corrigir problemas sistematicamente. O estudo também discute a aplicação de standards internacionais, sublinhando que a adesão a esses referenciais pode facilitar a

interoperabilidade e reduzir ambiguidades semânticas entre sistemas. Por fim, os autores defendem a necessidade de soluções que combinem validação em tempo real com avaliação pós-processamento, suportadas por técnicas como *data fusion*, *machine learning* e mecanismos de controlo distribuído.

No âmbito da análise de *Big Data* aplicada a fábricas inteligentes, Liu et al. (16) analisam de forma sistemática desafios e abordagens metodológicas, interligando diretamente os problemas de qualidade dos dados às suas implicações na eficácia dos sistemas analíticos e nos processos produtivos. Estes problemas afetam negativamente tarefas como controlo de qualidade, deteção precoce de falhas, manutenção preditiva e otimização de processos, levando os autores a defenderem que a qualidade dos dados é uma pré-condição para a maturidade analítica industrial. Liu et al. (16) estruturam a sua análise em torno de quatro dimensões principais da qualidade de dados, que se revelam recorrentes em contextos industriais (acurácia, completude, consistência e atualidade). Estas dimensões são posicionadas como elementos críticos na fiabilidade das análises, sendo que dados de má qualidade tendem a degradar o desempenho de modelos de *machine learning*, aumentar falsos positivos em sistemas de deteção de anomalias e induzir decisões operacionais incorretas. O artigo recomenda a integração de mecanismos de avaliação e correção da qualidade de dados no *pipeline* analítico, desde o momento da aquisição dos dados dos sensores até ao processamento em motores de análise.

Complementarmente, Al-Zaidawi e Çevik (21) propõem uma abordagem inovadora explorando o potencial da IA na monitorização da qualidade dos dados em redes IdC, focando-se na melhoria da qualidade dos dados através da aplicação de modelos avançados de aprendizagem profunda otimizados por algoritmos híbridos, permitindo acelerar a convergência, aumentar a precisão e melhorar a capacidade de deteção de padrões em fluxos de dados sensoriais. Esta linha de investigação aponta para a crescente relevância de métodos inteligentes na supervisão da integridade de fluxos em tempo real.

Estes contributos evidenciam que a qualidade dos dados constitui um desafio multidimensional, intrinsecamente ligado à maioria das iniciativas de digitalização industrial. Apesar da diversidade de abordagens propostas na literatura, verifica-se um consenso crescente quanto à necessidade de desenvolver sistemas de monitorização da qualidade dos dados que sejam simultaneamente automáticos, escaláveis e adaptáveis às especificidades dos contextos industriais. Embora já existam metodologias e métricas bem estabelecidas, a sua aplicação prática em ambientes produtivos reais permanece limitada, muitas vezes devido a barreiras tecnológicas ou organizacionais. Neste cenário, a incorporação sistemática de técnicas como o *data profiling* e a avaliação baseada em métricas objetivas torna-se cada vez mais crítica para garantir a fiabilidade dos dados, apoiar decisões informadas e potenciar a eficiência dos processos na I4.0.

As secções seguintes são baseadas nos artigos ***Data Quality Assessment in Smart Manufacturing: A Review*** (1) e ***Extensible Data Ingestion System for Industry 4.0*** (8).

2.1 Qualidade dos Dados

Com o rápido crescimento dos dispositivos IdC e o aumento da quantidade de dados que eles geram, garantir a qualidade desses dados tornou-se um desafio fundamental. Existem duas razões principais para a persistência dos problemas de qualidade dos dados em

sistemas IdC (9). Em primeiro lugar, as leituras dos sensores são frequentemente incompletas ou corrompidas por fatores imprevisíveis, como interferência eletromagnética, perda de pacotes ou problemas de processamento de sinal. Em segundo lugar, os dados muitas vezes percorrem longas distâncias, o que pode introduzir erros adicionais, latência e inconsistências. A maioria dos problemas de qualidade dos dados nestes cenários está relacionada à relação sinal-ruído (9), dificultando a extração de informações precisas a partir dos dados brutos. Todos estes fatores resultam numa grande diversidade de erros, como anomalias, valores em falta, desvios, ruído, valores constantes, preso no zero/falha, bias, duplicados, descontinuidades e inconsistências (20; 9). Quando se trata de múltiplas fontes de dados, os problemas são agravados por variações nos intervalos de recolha, diferenças nas unidades de medição, divergências nas especificações e a presença de incertezas inerentes (20). Os tipos mais comuns de erros de qualidade em dados provenientes de sensores são os valores anormais e os valores em falta (22).

Além disso, como mencionado por Mahanti (23), os problemas de qualidade não estão limitados a uma fase ou componente específico, podendo surgir em qualquer ponto do ciclo de vida dos dados, desde a sua criação ou aquisição, passando pela recolha via sensores, centros de contacto ou aplicações *web*, até ao processamento, armazenamento, transferência e eventual preservação dos dados. Problemas como erros de definição, incoerências no conteúdo ou falhas de apresentação podem ser introduzidos em qualquer uma destas fases, afetando diretamente a qualidade final da informação.

A qualidade dos dados tem sido amplamente reconhecida como um fator crítico em sistemas orientados por dados, sobretudo em ambientes industriais onde se exige tomada de decisão em tempo real (1). No entanto, a noção de "qualidade" é intrinsecamente multidimensional (24), assumindo significados distintos consoante o domínio de aplicação, os requisitos operacionais e as expectativas de cada organização (23). Esta variedade de interpretações torna difícil a adoção de uma definição única e consensual, sendo comum considerá-la como o grau em que os dados satisfazem requisitos explícitos e implícitos num determinado contexto de uso (25). Assim, a qualidade dos dados não pode ser reduzida a uma métrica única, devendo ser avaliada com base nas necessidades específicas dos utilizadores e nos objetivos do sistema onde se insere (3).

O conceito de qualidade dos dados é geralmente descrito através de dimensões, que representam atributos específicos e observáveis da qualidade dos dados. Estas dimensões fornecem uma base formal para avaliar, quantificar e gerir a qualidade de forma contínua (23; 26). Cada dimensão descreve um aspeto, e, quando medida adequadamente, oferece *insights* sobre a qualidade dos dados (20). A identificação destas dimensões é o primeiro passo para uma avaliação objetiva da qualidade, sendo também o ponto de partida para estratégias de melhoria contínua da informação (9; 27).

Neste contexto, diversos autores propuseram classificações distintas das DQD, agrupando-as segundo critérios funcionais, operacionais ou hierárquicos. Estas classificações ajudam a sistematizar a avaliação da qualidade em diferentes domínios e serão exploradas na secção seguinte.

2.2 Classificações para a Qualidade dos Dados

Ao longo dos anos, diversos autores propuseram diferentes taxonomias para organizar as DQD. Estas classificações consistem na organização das diversas DQD em grupos es-

pecíficos, cada um com foco em diferentes aspetos da sua avaliação. Esta abordagem permite não só uma compreensão mais aprofundada do papel de cada dimensão, como também facilita a sua aplicação prática em contextos distintos, ajustando-se às exigências específicas de cada domínio (1). A seguir, apresentam-se três das propostas presentes na literatura: Wang e Strong (26), Loshin (15) e Batini e Scannapieco (24).

2.2.1 Classificação de Wang e Strong (1996)

Wang e Strong (26) foram pioneiros ao estabelecer uma taxonomia de dimensões de qualidade baseada no ponto de vista dos utilizadores finais. Para eles, a qualidade dos dados deve ser definida segundo o critério da "adequação ao uso", ou seja, a medida em que os dados são úteis e apropriados para suportar as decisões e tarefas dos seus consumidores. Esta abordagem foi construída com base em métodos empíricos envolvendo centenas de participantes com experiência prática no uso de dados.

A sua classificação agrupa as dimensões em quatro categorias distintas: intrínseca, contextual, representacional e acessibilidade, conforme apresentado na Figura 2.1 de (1).

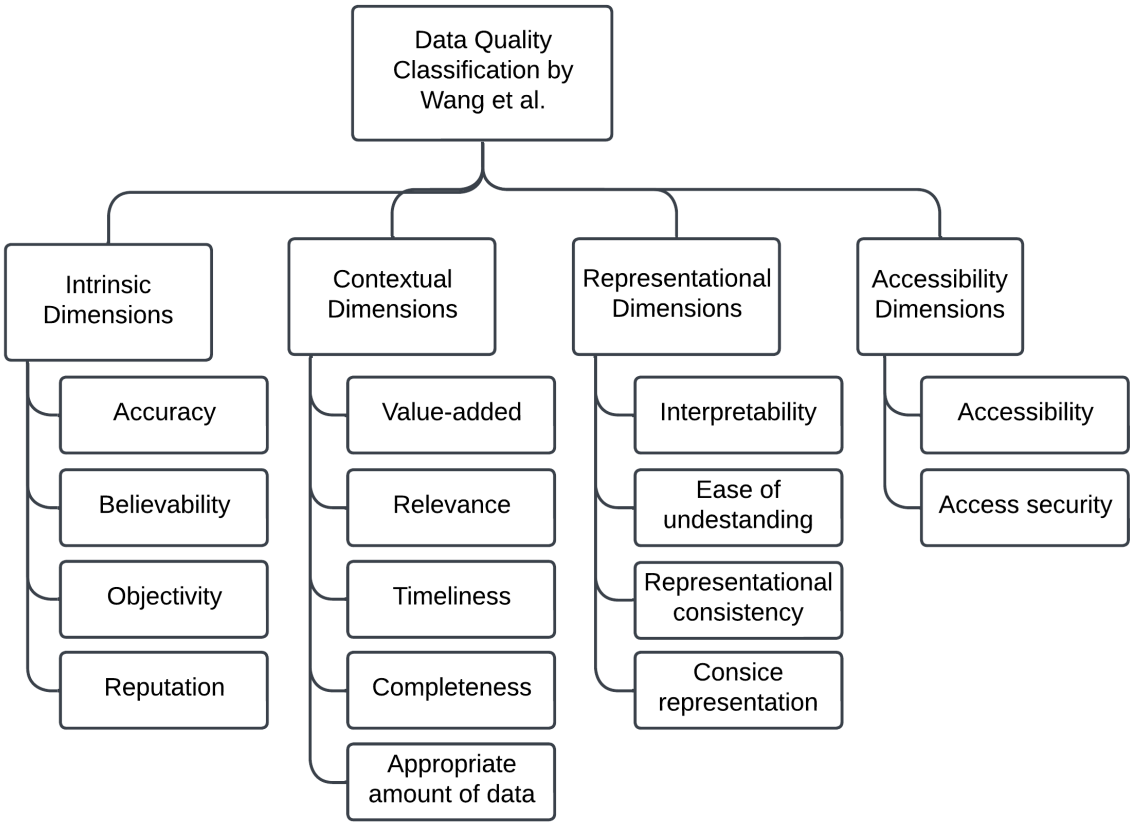


Figura 2.1: Classificação da qualidade dos dados por Wang e Strong. Fonte: (1)

O grupo da qualidade intrínseca dos dados refere-se aos atributos que os dados devem possuir por si mesmos, independentemente do contexto de uso e engloba as dimensões acurácia, credibilidade, objetividade e reputação. Estes atributos refletem a ideia de que dados devem ser corretos e fiáveis na sua essência, e não apenas úteis em determinadas aplicações. O grupo contextual representa o grau de adequação dos dados à tarefa específica para a qual são utilizados, ou seja, concentra-se principalmente no contexto da

tarefa, em vez do contexto da representação dos dados. Garantir uma elevada qualidade contextual pode ser desafiador, uma vez que as necessidades de negócio e os requisitos operacionais tendem a evoluir ao longo do tempo. As dimensões associadas incluem: valor acrescentado, relevância, atualidade, completude e quantidade apropriada de dados. A qualidade representacional está relacionada com a forma como os dados são apresentados e interpretados pelos utilizadores. Exige que os dados sejam claros, consistentes e concisos, de modo a facilitar a sua compreensão e análise. Este grupo inclui as dimensões de interpretabilidade, facilidade de compreensão, consistência representacional e apresentação concisa. Por fim, a qualidade de acessibilidade refere-se à disponibilidade e segurança no acesso aos dados por parte dos utilizadores. Esta dimensão torna-se especialmente relevante no contexto da digitalização e da adoção de sistemas distribuídos, nos quais o acesso eficiente e controlado aos dados é fundamental. Integra as dimensões de acessibilidade e segurança de acesso.

Esta taxonomia é uma das mais amplamente utilizadas na literatura e destaca-se por alinhar a qualidade dos dados com as necessidades dos utilizadores finais.

2.2.2 Classificação de Loshin (2011)

Loshin (15) propõe uma abordagem distinta para a avaliação da qualidade dos dados, estruturando as suas dimensões em três níveis hierárquicos, com o objetivo de facilitar a sua aplicação em diferentes contextos organizacionais. A Figura 2.2 ilustra a divisão proposta.

As dimensões são agrupadas em três categorias principais: dimensões intrínsecas, dimensões contextuais e dimensões qualitativas. As dimensões intrínsecas situam-se no nível mais básico e estão associadas à própria natureza dos dados, referindo-se às propriedades inerentes dos dados, ou seja, características independentes de qualquer associação com registos ou entidades específicas. Este grupo inclui dimensões estruturais e semânticas dos dados, como a acurácia, linhagem, consistência semântica e consistência estrutural. Estas dimensões descrevem os dados tal como são armazenados e representados, incluindo formatos, domínios e significados. No segundo nível estão as dimensões contextuais, que dependem do contexto de utilização dos dados e avaliam os dados em relação à sua utilização operacional e às regras comerciais em vigor. Estas dimensões medem a validade e coerência dos dados em relação a outros, sendo frequentemente condicionadas por políticas internas e processos automatizados. As dimensões deste grupo são a completude, consistência, *currency*, atualidade, razoabilidade e identificabilidade. Por fim, as dimensões qualitativas representam o nível mais elevado e abrangente, avaliando perceções ou requisitos subjetivos. Estas dimensões oferecem uma visão integrada e subjetiva da qualidade, avaliando a medida em que os dados satisfazem as expectativas dos utilizadores e se alinham com os objetivos estratégicos da organização. Funcionam como uma síntese das dimensões intrínsecas e contextuais, sendo particularmente úteis em contextos onde métricas quantitativas claras não são facilmente aplicáveis. Estas dimensões permitem uma avaliação holística da qualidade dos dados, focando-se no valor global da informação para o negócio.

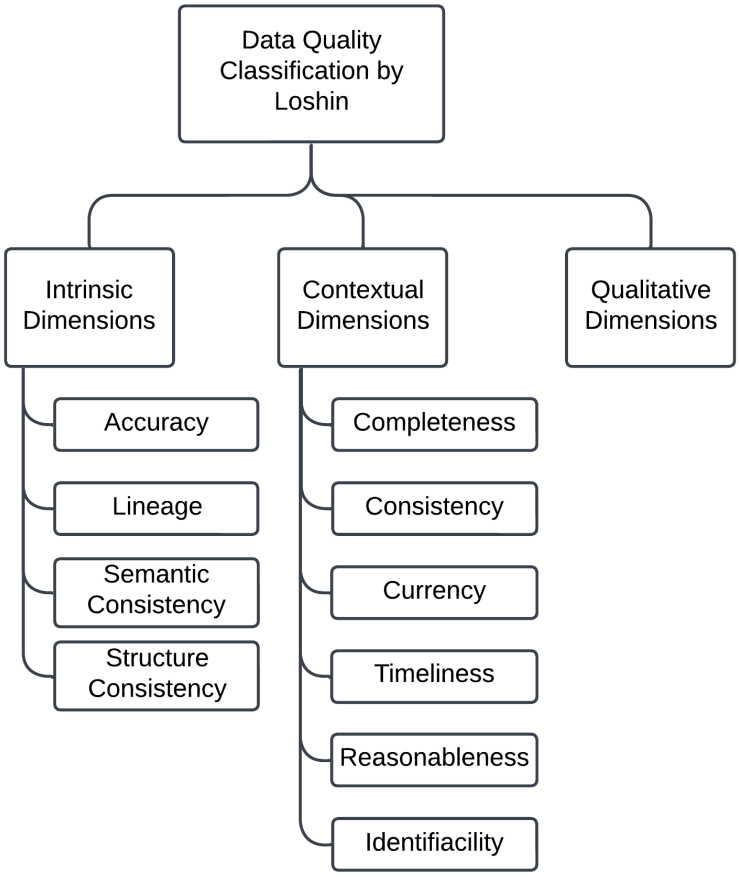


Figura 2.2: Classificação da qualidade dos dados por Loshin. Fonte: (1)

2.2.3 Classificação de Batini e Scannapieco (2016)

Batini e Scannapieco (24) apresentam uma das classificações mais detalhadas das DQD, organizando-as segundo uma perspetiva funcional. A sua proposta visa facilitar a identificação, avaliação e monitorização das dimensões relevantes em diferentes fases do ciclo de vida dos dados, agrupando dimensões semelhantes segundo a sua função conceptual. A classificação está estruturada em oito grupos, tal como pode se ver na Figura 2.3 de (1). Esta abordagem permite comparar dimensões associadas a diferentes tipos de informação (por exemplo, dados estruturados, imagens, dados *web*).

Os oito grupos identificados são descritos de forma clara e abrangem tanto dimensões clássicas como outras mais ligadas à perceção do utilizador e à confiança nas fontes. O grupo da acurácia está relacionado com o grau de fidelidade dos dados face à realidade que representam. Este grupo inclui dimensões como acurácia, correção, validade e precisão. A acurácia é ainda subdividida em dois componentes: acurácia estrutural, que se desdobra em acurácia sintática (avaliando a distância entre um valor e os elementos do seu domínio de definição) e acurácia semântica (proximidade do valor ao valor verdadeiro) e acurácia temporal, que mede a rapidez com que as atualizações nos dados refletem mudanças reais. O grupo da completude abrange as dimensões de completude, pertinência e relevância, avaliando se os dados têm amplitude, profundidade e alcance suficientes para a tarefa em questão. O grupo da redundância inclui as dimensões redundância, minimalidade, compacidade e concisão, e está relacionado com a eficiência da representação da informação.

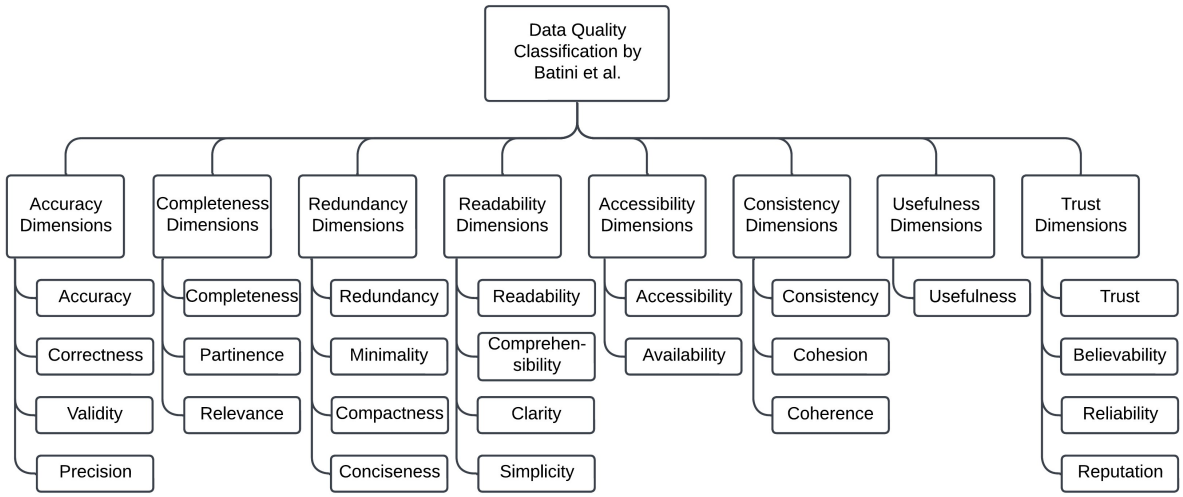


Figura 2.3: Classificação da qualidade dos dados por Batini e Scannapieco. Fonte: (1)

As dimensões de legibilidade dizem respeito à facilidade com que a informação pode ser interpretada pelos seus utilizadores. Este grupo inclui legibilidade, compreensibilidade, clareza e simplicidade, promovendo uma melhor utilização da informação. O grupo da acessibilidade inclui as dimensões acessibilidade e disponibilidade, centrando-se na capacidade dos utilizadores de aceder à informação, tendo em conta limitações culturais, físicas ou tecnológicas. O grupo da consistência avalia a conformidade da informação com regras formais e semânticas, como regras de integridade, regras de negócio e restrições estruturais. Engloba as dimensões consistência, coesão e coerência. O grupo da utilidade inclui apenas a dimensão utilidade, que se relaciona com o valor prático da informação e o benefício direto que o utilizador retira da sua utilização. De acordo com este princípio, devem ser recolhidos apenas os dados que sejam úteis para o propósito definido, seja ele operacional ou secundário, garantindo que a informação é tratada exclusivamente para fins específicos, explícitos e legítimos, e não reutilizada de forma incompatível com esses objetivos. Por fim, o grupo da confiança agrega as dimensões confiança, credibilidade, confiabilidade e reputação, avaliando o grau de fidedignidade atribuído à informação, incluindo aspetos como segurança, autenticidade e proveniência. Esta estrutura de agrupamento oferece um modelo conceptual robusto para a avaliação da qualidade dos dados em domínios heterogéneos, promovendo a interoperabilidade dos critérios de qualidade entre contextos técnicos, operacionais e organizacionais.

Entre os vários fatores que comprometem a qualidade dos dados, destaca-se a presença de valores atípicos e anomalias, que requerem atenção especial. A próxima secção explora este problema em pormenor, antes de se avançar para a análise das DQD.

2.3 Deteção de Anomalias

A deteção de anomalias é uma componente essencial na garantia da qualidade dos dados em ambientes industriais, especialmente no contexto da IdC. Dados provenientes de sensores podem conter erros que comprometem o seu valor informativo, conduzindo a decisões incorretas ou falhas no desempenho dos sistemas (22). Entre os principais problemas de qualidade dos dados neste domínio, destacam-se os valores ausentes e os *outliers* (22), cuja presença compromete significativamente a fiabilidade das análises e dos processos

de decisão automática. Assim, a identificação precoce de comportamentos inesperados é crítica para garantir a eficiência dos sistemas industriais modernos.

O termo *outlier*, também conhecido como anomalia, tem origem na estatística (28) e refere-se a observações que se desviam significativamente do padrão geral dos dados (29). De forma geral, uma anomalia pode ser entendida como um padrão nos dados que não está em conformidade com uma definição bem estabelecida de comportamento normal (30). Segundo Ahmad e Purdy (31), uma anomalia corresponde a um ponto no tempo em que o comportamento do sistema se revela invulgar face ao seu histórico. No entanto, esse comportamento atípico não implica necessariamente um problema. Uma variação pode resultar de uma situação negativa, como um aumento anómalo da temperatura de um motor que indica falha iminente, ou de uma situação positiva, como um pico de cliques numa página *web* de um produto recém-lançado, o que sinaliza uma elevada procura. Em ambos os casos, os dados são invulgares e podem justificar uma ação (31).

A deteção de anomalias tem sido amplamente investigada em diversas áreas (32), como na manutenção preditiva, na deteção de fraudes, na monitorização e análise de falhas, entre outras (31). Trata-se, contudo, de um problema complexo que exige uma compreensão profunda da distribuição dos dados e da identificação de padrões inconsistentes (33). Essa complexidade é agravada por dois obstáculos principais: o elevado volume de dados e a natureza variada dos mesmos (34).

Em particular, a deteção de anomalias apresenta desafios adicionais (35). A diversidade de padrões anómalos, a necessidade de escalar o processamento a grandes volumes de dados em tempo real e a minimização de sobrecarga computacional em ambientes distribuídos são apenas alguns exemplos (35). Além disso, a continuidade temporal desempenha um papel central: alterações abruptas, padrões invulgares e eventos inesperados devem ser analisados com base no seu contexto temporal (35).

Segundo Chatterjee e Ahmed (32), o problema da deteção de anomalias em contextos IdC pode ser categorizado segundo quatro perspetivas principais. Estas categorias referem-se à abordagem adotada na formulação do problema, ao tipo de aplicação, ao tipo de anomalia considerada e à latência dos algoritmos envolvidos.

Do ponto de vista metodológico (32), distinguem-se três grandes grupos: abordagens geométricas, métodos estatísticos e modelos de aprendizagem automática e profunda. As abordagens geométricas baseiam-se em estratégias de distância e densidade, assumindo que os dados normais e anómalos ocupam regiões distintas do espaço de características. A classificação de um ponto como anómalo depende de um limiar aplicado à distância estimada em relação ao conjunto de dados. Já os métodos estatísticos procuram modelar o comportamento normal dos dados através de distribuições matemáticas ou abordagens preditivas, considerando anómalos os valores que se desviam significativamente do comportamento esperado. Por sua vez, os modelos baseados em aprendizagem automática e aprendizagem profunda demonstram melhor adaptação a dados complexos. A escolha do modelo depende da natureza dos dados: por exemplo, *Long Short-Term Memory* (LSTM) e modelos transformadores são adequados a dados sequenciais, como séries temporais, enquanto codificadores automáticos e redes neuronais convolucionais apresentam melhor desempenho com dados não sequenciais. Estes modelos procuram estabelecer fronteiras de decisão para distinguir comportamentos normais de comportamentos anómalos, podendo operar sob regimes supervisionado, não supervisionado ou semi-supervisionado, consoante a disponibilidade de rótulos nos dados de treino.

Quanto à aplicação (32), as técnicas podem ser agrupadas em aplicações construtivas, destrutivas e de limpeza de dados. Aplicações construtivas incluem, por exemplo, a monitorização do comportamento de idosos para prevenção de quedas ou a monitorização de processos agrícolas. Por contraste, aplicações destrutivas referem-se à deteção de ciberataques e manipulação maliciosa de sistemas críticos, como ilustrado em diversos trabalhos sobre segurança em IdC. Por fim, as aplicações de limpeza de dados visam eliminar ruído e picos espúrios em sinais sensoriais, contribuindo para a melhoria da qualidade dos dados antes da sua análise.

A categorização das anomalias (32) pode incluir anomalias pontuais, que correspondem a instâncias isoladas que se desviam significativamente do comportamento esperado. São o tipo mais simples de anomalia e podem ser identificadas isoladamente, sem necessidade de informação contextual. Já as anomalias contextuais surgem quando determinado valor é considerado anómalo apenas no contexto em que ocorre. Um exemplo clássico é uma temperatura baixa que pode ser aceitável no inverno, mas suspeita no verão. Para tal, é necessário que o conjunto de dados inclua atributos que permitam estabelecer este contexto, como localização espacial, marca temporal ou outras variáveis ambientais. Por sua vez, as anomalias coletivas resultam da ocorrência simultânea de vários valores que, em conjunto, formam um padrão anómalo, mesmo que individualmente não o sejam. Estas só podem ser detetadas em conjuntos de dados com instâncias relacionadas entre si, como séries temporais ou dados espaciais. Uma anomalia pontual ou uma anomalia coletiva também pode ser considerada uma anomalia contextual, caso seja analisada em função do contexto. Assim, um problema de deteção de anomalias pontuais ou coletivas pode ser reformulado como um problema de deteção de anomalias contextuais, desde que se incorpore informação contextual adequada (30).

Em termos de latência (32), distinguem-se algoritmos *online* e *offline*. Os primeiros operam em tempo real, processando os dados à medida que são gerados, e incluem abordagens baseadas em janelas deslizantes ou estimativas incrementais. Já os algoritmos *offline* assumem acesso ao conjunto completo de dados e utilizam técnicas mais sofisticadas, frequentemente com maior custo computacional. Modelos como LSTM ou o classificador *Naive Bayes* são, por vezes, treinados em modo *offline* e posteriormente aplicados em tempo real em cenários de inferência contínua.

Além destas classificações, as abordagens de deteção de anomalias podem também ser agrupadas em métodos supervisionados e não supervisionados (36). A escolha entre uma abordagem ou outra depende de fatores como a disponibilidade de rótulos nos dados, o tipo de anomalia a ser detetada, o volume de dados e a necessidade de processamento em tempo real.

Outro aspeto relevante nas técnicas de deteção de anomalias é a forma como os resultados são apresentados (30). Algumas abordagens atribuem uma pontuação de anomalia a cada instância, refletindo o grau de desvio em relação ao comportamento esperado. Este tipo de saída permite aos analistas definir limiares adaptados ao domínio para selecionar os casos mais relevantes. Outras técnicas atribuem diretamente rótulos binários (por exemplo, anómalo ou normal), simplificando a análise mas reduzindo a flexibilidade na interpretação. A escolha entre pontuações contínuas e classificações binárias depende do objetivo da monitorização, dos requisitos de interpretabilidade e da criticidade das decisões associadas.

Em séries temporais industriais, diversas abordagens têm sido utilizadas para a deteção

de anomalias, cada uma com características que favorecem determinados contextos operacionais.

Uma técnica probabilística particularmente eficiente para a deteção de anomalias e baseada em ordenação é o *t-digest*, um algoritmo concebido para gerar esboços compactos de dados que permitem calcular, com elevada precisão, quantis, medianas e médias (37). Esta estrutura de dados consiste no agrupamento de amostras com valores reais, onde o tamanho de cada grupo é regulado por uma função de escala, diferenciando-se assim dos métodos de agrupamento tradicionais (38; 39).

O *t-digest* apresenta duas propriedades fundamentais que o tornam particularmente adequado para aplicações em ambientes industriais: elevada precisão na estimativa de quantis, mesmo em distribuições arbitrárias, e baixos requisitos de memória, com suporte para operação *online* e limites de memória constantes (37; 39). Além disso, os esboços gerados podem ser combinados, permitindo a consolidação de resumos de dados provenientes de diferentes fontes ou janelas temporais (38).

A estrutura do *t-digest* é especialmente útil em contextos de monitorização contínua e deteção precoce de anomalias, sendo frequentemente utilizada como etapa final no processamento de séries temporais (37; 40). Nestes casos, os limites de deteção podem ser definidos com base em quantis históricos estimados, ajustando-se dinamicamente ao comportamento do processo (37).

Os benefícios do *t-digest* têm sido particularmente evidentes em contextos industriais, onde a eficiência na monitorização pode traduzir-se diretamente em ganhos económicos, pela prevenção de falhas e maximização do tempo de atividade dos sistemas (37). Em contraste, em ambientes de investigação, a sua adoção é menos comum, dado que os conjuntos de dados tendem a ser de menor dimensão e os requisitos de desempenho computacional são, geralmente, menos exigentes (37).

Outra abordagem amplamente utilizada na análise de séries temporais é a decomposição da tendência sazonal com recurso ao método *Seasonal-Trend decomposition using Loess* (STL), que permite separar uma série em três componentes distintos: tendência, sazonalidade e residual (41). Esta decomposição facilita a deteção de variações anómalas ao isolar o componente residual, onde se concentram os desvios não explicados pelos padrões regulares do processo (42). A elevada interpretabilidade dos resultados torna o STL particularmente atrativo em contextos que exigem transparência analítica (42).

O STL recorre à suavização local com *Loess* para capturar padrões sazonais e tendências de longo prazo de forma flexível (41). A sua estrutura modular assenta em dois ciclos principais: um ciclo interno, responsável pela suavização da sazonalidade e da tendência, e um ciclo externo, que regula a robustez do modelo através da reponderação dos resíduos (43). Esta arquitetura torna o algoritmo suficientemente robusto para lidar com variações estruturais e ruído nos dados, mantendo simultaneamente a simplicidade de implementação e baixos custos computacionais (41).

Além da sua eficácia na deteção de anomalias, o método STL tem demonstrado um desempenho consistente em tarefas de previsão, podendo ser integrado com técnicas como o suavizamento exponencial para melhorar a acurácia preditiva (43). A sua capacidade de decompor dados com padrões complexos reforça a sua aplicabilidade em múltiplos domínios industriais.

Adicionalmente, os modelos baseados em redes neuronais, como as LSTM, têm ganho destaque na deteção de anomalias em séries temporais complexas, devido à sua capacidade de capturar padrões não lineares e dependências de longo prazo (44; 45). Ao contrário das redes neuronais tradicionais, que não conseguem lidar eficazmente com sequências temporais dependentes de contexto, as redes neuronais recorrentes introduzem mecanismos de retroalimentação que permitem considerar o estado anterior como entrada adicional (43). No entanto, as redes neuronais recorrentes clássicas enfrentam limitações como o problema do desaparecimento do gradiente, que compromete a aprendizagem de dependências distantes no tempo (43).

As redes LSTM surgem como uma extensão das redes neuronais recorrentes concebida para ultrapassar essas limitações, integrando um mecanismo de memória interna que permite preservar informação relevante ao longo do tempo (43). A estrutura da LSTM assenta num "estado da célula", cuja atualização é controlada por três portas: a porta de esquecimento, que decide quais as informações antigas a descartar, a porta de entrada, que determina quais as novas informações a incorporar e a porta de saída, que regula o que deve ser transferido para a próxima etapa (43; 44).

Este tipo de arquitetura provou ser particularmente eficaz na modelação e previsão de séries temporais em ambientes industriais, permitindo a extração automática de características abstratas a partir de dados brutos (45). A sua robustez na identificação de padrões dinâmicos e complexos torna as LSTM uma escolha relevante na deteção de anomalias em contextos com elevada variabilidade e requisitos de sensibilidade preditiva.

A deteção eficaz de anomalias em dados industriais não só contribui para a fiabilidade dos sistemas de monitorização e controlo, como também se relaciona diretamente com várias dimensões da qualidade dos dados, como a acurácia, consistência e atualidade. Assim, este processo deve ser encarado como parte integrante de qualquer estratégia de gestão da qualidade de dados, particularmente em contextos de produção intensiva e sensorização contínua.

2.4 Dimensões da Qualidade dos Dados

As DQD constituem atributos ou características fundamentais que determinam se os dados são adequados para o uso a que se destinam (15). Estas dimensões permitem avaliar a qualidade de forma objetiva, ao estabelecer critérios mensuráveis que orientam tanto a avaliação como a melhoria contínua dos dados (1). Segundo Wang e Strong (26), uma dimensão representa um conjunto de atributos que refletem um único aspeto da qualidade dos dados, funcionando como uma unidade conceptual que facilita a sua gestão e controlo.

As DQD podem ser vistas sob duas perspetivas complementares: a extensão (valores dos dados) e a intensão (esquemas e estruturas) (24). Embora frequentemente definidas de forma qualitativa, referindo-se às propriedades em geral, estas dimensões carecem de métricas explícitas, sendo necessário associar indicadores quantitativos específicos para viabilizar a sua monitorização e análise (24).

Para garantir o retorno sobre o investimento em iniciativas de gestão da qualidade dos dados, é essencial priorizar dimensões que sejam pertinentes e alinhadas com os objetivos organizacionais (23). Medir todas as dimensões possíveis pode proporcionar uma visão mais abrangente, mas, na prática, é comum focar-se nas mais impactantes para o processo

de decisão. Assim, a correta identificação e aplicação das DQD não só permite detetar lacunas e oportunidades de melhoria ao longo do fluxo de informação, como também assegura a conformidade com os requisitos operacionais (15; 27).

A literatura apresenta uma variedade extensa de dimensões propostas, refletindo a complexidade e a diversidade dos contextos em que os dados são utilizados (27). Contudo, em ambientes industriais e sistemas de IdC, certas dimensões assumem maior relevância devido às exigências operacionais em tempo real. Nestes tipos de ambientes, os principais problemas da qualidade dos dados relacionam-se diretamente com quatro dimensões, a acurácia, a completude, a consistência e a atualidade (16). Estas dimensões permitem caracterizar e quantificar aspetos específicos da informação, como a acurácia dos valores registados, a presença de todos os dados esperados, a coerência entre dados relacionados e a atualidade dos dados. A Tabela 2.1, baseada em (1), representa a relação entre estas dimensões e os principais problemas associados à qualidade dos dados em ambientes industriais, conforme discutido por (15).

Tabela 2.1: Relação entre as quatro dimensões e os principais problemas em ambientes industriais.

Dimensão	**Problema**	**Descrição**
Acurácia	Anomalias, *outliers*, ruído e dados inválidos.	Os dados desviam-se dos padrões normais e encontram-se fora do conjunto de valores plausíveis.
Completude	Dados ou valores em falta.	Existem valores nulos ou ausentes nos registos de dados.
Consistência	Inconsistências, redundâncias e duplicação de dados.	Os dados provenientes de diferentes fontes são contraditórios, ou os dados da mesma observação apresentam incoerências.
Atualidade	Dados antigos, desatualizados ou com problemas de alinhamento temporal.	Os dados estão desatualizados e não refletem o estado atual do sistema.

2.4.1 Acurácia

A acurácia é uma das dimensões fundamentais da qualidade dos dados, particularmente em ambientes industriais e sistemas de IdC, representando o grau em que os dados refletem corretamente a realidade que pretendem descrever. De acordo com Loshin (15), a acurácia refere-se à medida em que os valores dos dados representam informação correta e fiável, sendo essencial para a tomada de decisão baseada em dados. Mahanti (23) define a acurácia como o grau em que cada dado descreve corretamente um objeto do mundo real. Propõe a sua medição em dois níveis: ao nível do registo, por exemplo, se um conjunto de leituras de sensores reflete com precisão o estado de uma máquina, e ao nível do elemento de dado, como no caso de uma leitura de temperatura que corresponde ou não ao valor real. De forma geral, esta dimensão avalia a proximidade dos dados ao valor verdadeiro (11), embora, na prática, o valor real muitas vezes seja desconhecido ou de difícil acesso

(46).

Adicionalmente, a acurácia pode ser entendida como o grau em que os dados armazenados num sistema refletem fielmente os objetos, entidades ou eventos do mundo real. Neste sentido, Mahanti (23) propõe três questões-chave para a avaliação da acurácia:

- Os dados caracterizam com precisão os valores do mundo real que devem representar e as entidades da vida real que modelam?
- Até que ponto os dados capturam corretamente o que foram concebidos para capturar?
- Os dados representam com precisão a realidade ou uma fonte verificável?

Batini e Scannapieco (24) apresentam uma distinção entre três tipos de acurácia: sintática, que diz respeito à conformidade com o domínio de definição, como o tipo ou formato dos dados, semântica, que avalia se os dados fazem sentido dentro do contexto da linguagem ou do sistema, e temporal, que considera a atualidade, estabilidade e frequência de atualização dos dados. Goknil et al. (9) também descrevem a acurácia como o grau de precisão com que os dados armazenados refletem a realidade. Tverdal et al. (11) reforçam esta perspetiva, apontando que a acurácia corresponde ao grau de semelhança de uma quantidade medida com o seu valor real.

Nos ambientes industriais e sistemas de IdC, Liu et al. (16) destacam que a falta de acurácia pode manifestar-se através de valores inválidos, ruído, *outliers*, leituras incorretas e dados defeituosos, sendo estas situações frequentemente causadas por falhas de equipamentos, erros humanos ou condições ambientais adversas. Karkoucha et al. (47) introduzem uma abordagem quantitativa à acurácia em sensores, definindo-a como o erro sistemático absoluto máximo admissível. Este é representado por um intervalo simétrico $[v - \alpha, v + \alpha]$, dentro do qual se espera que o valor real se encontre. O valor de α representa o erro permitido com base nas características técnicas do sensor, como a sua classe de precisão ou gama de medição. Por outro lado, Kim e Lee (46) salientam que, na ausência de um valor de referência absoluto, é possível recorrer à utilização de estimativas estatísticas de intervalos de confiança para inferir a acurácia dos dados gerados por sensores. Neste contexto, um valor pode ser considerado preciso se se situar dentro do intervalo de confiança estabelecido (46).

Assim, a acurácia assume um papel essencial na integridade dos sistemas de monitorização e controlo, devendo ser tratada com particular atenção durante o desenho e validação de soluções orientadas por dados.

2.4.2 Completude

A completude é uma das dimensões mais básicas e essenciais na avaliação da qualidade dos dados (23), representando o grau em que a informação necessária está presente e disponível para utilização. Segundo Mahanti (23), a completude refere-se ao grau em que todos os valores esperados se encontram presentes no conjunto de dados, estando relacionada com a existência de valores em falta ou nulos. A ausência de informação pode resultar em análises incompletas, afetando negativamente a fiabilidade dos sistemas analíticos (16).

Para avaliar a completude de forma estruturada, Mahanti (23) propõe a sua análise em

três níveis: elemento do dado, registo e conjunto de dados. Esta abordagem é especialmente útil em contextos industriais, onde a informação provém de diversas fontes (1). Ao nível do elemento do dado, importa distinguir entre atributos obrigatórios, que devem sempre conter um valor, e atributos inaplicáveis, que só assumem valor em determinadas condições. Ao nível do registo, verifica-se se todos os campos obrigatórios estão preenchidos. Já ao nível do conjunto de dados, avalia-se se todos os registos esperados estão presentes, embora este nível possa ser insuficiente em cenários de elevada granularidade, como nos sistemas de IdC.

Para além disso, torna-se fundamental compreender como os dados ausentes são representados. Em muitos casos, espaços em branco, valores como "NA", "não aplicável" ou "desconhecido" são também indicadores de incompletude (23). Para apoiar a avaliação da completude, Mahanti (23) propõe ainda um conjunto de questões orientadoras:

- Toda a informação necessária está disponível?
- Existem dados críticos em falta nos registos?
- Todos os conjuntos de dados estão registados?
- Todos os itens de dados obrigatórios foram preenchidos?

À semelhança da abordagem de Mahanti, também Batini e Scannapieco (24) propõem uma avaliação da completude em três níveis distintos: completude do esquema, completude da coluna e completude populacional. A completude do esquema refere-se ao grau em que os conceitos e propriedades esperados estão presentes no modelo de dados, sendo particularmente relevante na fase de conceção e definição da estrutura da base de dados. A completude da coluna avalia a proporção de valores existentes numa determinada coluna ou atributo, ou seja, verifica a existência de valores ausentes em campos específicos. Por fim, a completude populacional mede se o conjunto de dados contém todos os registos esperados em relação a uma população ou universo de referência. Esta caracterização é especialmente adequada ao modelo relacional, onde a estrutura dos dados permite uma análise sistemática da presença ou ausência de informação.

Loshin (15) destaca que a completude está relacionada com a expectativa de que certos atributos tenham valores atribuídos. Kim e Lee (46) definem a completude como o grau em que a informação de contexto está presente. Além disso, propõem a distinção entre a informação gerada por sensores individuais e a informação agregada no sistema.

Num cenário mais técnico, Karkoucha et al. (47) propõem basear-se na proporção de elementos não interpolados sobre o total de elementos (interpolados e não interpolados) numa janela de tempo específica. Esta abordagem é útil na análise de fluxos de dados contínuos, em que falhas esporádicas na recolha de dados podem comprometer a completude global. Os autores alertam ainda que, mesmo com mecanismos de confirmação e recuperação de pacotes perdidos, os dados podem já estar obsoletos quando finalmente recuperados, reduzindo a utilidade da informação. Tverdal et al. (11) definem a completude como o grau em que todas as partes dos dados estão disponíveis, sem informação em falta, enquanto Cichy e Rass (27) realçam que esta dimensão reflete a suficiência da informação em termos de abrangência, profundidade e âmbito para a tarefa em questão.

Assim, a completude é particularmente crítica em contextos onde a ausência de dados essenciais compromete diretamente a capacidade de representação dos estados reais do sistema.

2.4.3 Consistência

A consistência é uma dimensão central da qualidade dos dados que se refere à uniformidade, coerência e sincronização da informação entre diferentes fontes, sistemas ou instâncias de aplicação (1). Segundo Mahanti (23), a consistência significa que os valores dos dados devem ser idênticos em todas as instâncias relevantes e que a sua formatação e apresentação devem ser uniformes em todo o conjunto de dados. Esta dimensão está particularmente relacionada com problemas como valores contraditórios entre fontes, duplicações e representações redundantes de um mesmo objeto (16). É importante salientar que consistência não implica necessariamente acurácia: dados consistentes podem ainda assim estar errados, mas dados inconsistentes indicam, com alta probabilidade, a presença de valores incorretos ou inválidos (1).

Para avaliar a consistência, Mahanti (23) propõe três níveis distintos: consistência ao nível do registo, consistência entre registos, e consistência ao nível do conjunto de dados. Esta estrutura permite identificar tanto contradições internas nos dados como discrepâncias entre diferentes representações de uma mesma entidade. Exemplos incluem: diferentes fontes que apresentam valores divergentes para o mesmo sensor, formatos incompatíveis de representação de datas ou unidades, e atributos relacionados que não refletem coerência entre si. A consistência também requer que os dados sejam armazenados de forma uniforme entre tabelas, documentos e sistemas. Batini e Scannapieco (24) reforçam que a consistência está associada à violação de regras semânticas definidas sobre os dados. Estas regras, como restrições de integridade ou dependências entre atributos, devem ser estabelecidas em colaboração com especialistas do domínio, assegurando que refletem corretamente as relações e restrições inerentes ao contexto da aplicação. A verificação da consistência, neste sentido, implica a validação de regras semânticas sobre os dados.

Loshin (15) define a consistência como a uniformidade da representação de dados entre diferentes fontes ou instâncias, destacando a necessidade de sincronização das definições, estruturas, significados e formatos dos dados em toda a organização. Cichy e Rass (27) complementam esta visão ao referirem que a consistência assegura a compatibilidade dos dados ao longo de múltiplos sistemas, evitando interpretações incorretas.

Para orientar a avaliação da consistência, Mahanti (23) propõe um conjunto de questões-chave:

- Existe uma representação única dos dados?
- Fontes distintas fornecem informação contraditória sobre o mesmo objeto?
- Que valores contradizem outros dentro ou entre conjuntos de dados?
- Os dados estão armazenados no mesmo formato em todas as tabelas e sistemas?
- Os valores são uniformes entre diferentes fontes ou ambientes?
- A relação entre atributos dependentes é logicamente coerente?

Desta forma, a consistência é fundamental para garantir que os dados utilizados em processos operacionais e analíticos representam de forma harmonizada e coerente a realidade que se pretende representar.

2.4.4 Atualidade

A atualidade é uma dimensão essencial da qualidade dos dados, especialmente em sistemas que exigem resposta em tempo real, como os ambientes industriais (1). Esta dimensão representa o grau em que os dados estão atualizados e disponíveis no momento certo para apoiar a tarefa em causa (16). Segundo Mahanti (23), a atualidade reflete o impacto temporal na utilidade da informação. A sua importância advém do facto de que dados que são tecnicamente atuais podem tornar-se inúteis se não estiverem disponíveis no momento em que são necessários. Por exemplo, um horário atualizado de aulas perde utilidade se só for disponibilizado após o início das aulas (24).

Loshin (15) define a atualidade como a expectativa de que os dados estejam disponíveis quando são esperados e necessários, enquanto Cichy e Rass (27) destacam a sua ligação com a disponibilidade e acessibilidade imediata da informação para apoiar decisões operacionais. Mahanti (23) reforça que os dados devem ser capturados o mais rapidamente possível após a ocorrência de um evento, sendo depois disponibilizados em tempo útil para responder às necessidades do negócio. Neste contexto, a atualidade é particularmente crítica para sensores e dispositivos em tempo real, onde atrasos na captura, transmissão ou processamento podem comprometer a fiabilidade do sistema.

Problemas comuns associados à atualidade incluem a presença de dados desatualizados, desalinhamentos temporais ou latência na atualização da informação (16). Estes problemas são especialmente relevantes em sistemas distribuídos ou em fluxos contínuos de dados, onde o tempo de chegada e de processamento pode variar significativamente. Karkoucha et al. (47) destacam que a diferença entre a marca temporal atual e a marca de registo pode ser usada para identificar atrasos críticos, que afetam diretamente a eficácia da análise.

Para apoiar a avaliação desta dimensão, Mahanti (23) propõe um conjunto de questões orientadoras:

- Qual é a diferença temporal entre a ocorrência do evento e a captura dos dados?
- Quanto tempo decorre até os dados estarem disponíveis para o utilizador final?
- Os dados estão acessíveis no momento em que são necessários?
- Estão atualizados para a tarefa em causa?
- A frequência de atualização é adequada aos requisitos do negócio?

Assim, a atualidade não se limita à presença de dados recentes, mas implica que estes estejam disponíveis em tempo útil para cumprir o seu propósito, sendo uma condição essencial para garantir decisões eficazes e responsivas em sistemas orientados por dados.

2.5 Métricas para a Qualidade dos Dados

Após a identificação das principais DQD, torna-se necessário estabelecer formas objetivas de quantificar cada uma destas dimensões. Neste contexto, surgem as métricas de qualidade dos dados, que permitem avaliar, de forma mensurável, o grau de conformidade dos dados com os requisitos operacionais e de negócio (23; 24).

As métricas de qualidade são instrumentos essenciais para a monitorização contínua, a

deteção de problemas, a comparação entre sistemas e a tomada de decisões baseada em evidências. Permitem identificar lacunas e oportunidades de melhoria ao longo do fluxo de informação, assegurando a conformidade com padrões técnicos e operacionais. De acordo com Batini e Scannapieco (24), a medição é indispensável para a formalização de qualquer estratégia de gestão da qualidade dos dados. Mahanti (23) reforça que a definição das métricas deve considerar o contexto de aplicação e os requisitos específicos de informação, especialmente em domínios como a I4.0, onde os dados são gerados em grande volume e com elevada volatilidade.

A avaliação da qualidade dos dados pode apoiar-se em métricas quantitativas, baseadas em dados observáveis, ou métricas qualitativas, suportadas por julgamentos subjetivos (23). Em contextos operacionais, como os sistemas de IdC, é preferível recorrer a métricas objetivas, pois estas reduzem o risco de interpretações erróneas (15). Estas métricas estão tipicamente associadas às dimensões que pretendem quantificar (23), podendo uma única dimensão ser avaliada por múltiplas métricas (20), cuja escolha depende dos objetivos analíticos e da natureza dos dados disponíveis. Diversos autores propõem a sua formalização através de expressões matemáticas (48), ou baseiam-se em observações recolhidas ao longo do tempo (49). As métricas podem assumir diferentes formas: valores normalizados entre 0 e 1, percentagens, frequências relativas, classificações binárias, ou estimativas probabilísticas (8; 9; 13; 50).

Cichy e Rass (27) salientam que as métricas devem ser aplicáveis em diferentes cenários, sensíveis a alterações reais nos dados e economicamente viáveis, sobretudo em ambientes com fluxos contínuos. Liu et al. (16) destacam que, em contextos industriais, as métricas devem ser granulares, adaptáveis a diversas fontes e robustas face a falhas como perda de pacotes ou desincronizações.

Antes de apresentar as métricas associadas a cada dimensão, importa compreender as características que tornam uma métrica de qualidade de dados eficaz. De acordo com Loshin (15), uma boa métrica deve apresentar oito características essenciais: deve ser clara e bem definida, mensurável, relevante para os objetivos do negócio e passível de controlo caso algum valor não seja o esperado. Além disso, deve permitir uma representação visual intuitiva, ser facilmente reportável, rastreável ao longo do tempo e suficientemente detalhada para identificar causas de variações no desempenho. Estas propriedades estão sintetizadas em (1) e ajudam a garantir que os indicadores definidos sejam interpretáveis, fiáveis e acionáveis.

A aplicação das métricas pode ocorrer em diferentes níveis de análise, proporcionando uma visão mais detalhada e adaptada aos objetivos da monitorização (23). Ao nível do registo, considera-se o conjunto completo de atributos de uma instância, permitindo uma visão holística. No nível do elemento de dado, cada atributo é analisado individualmente, o que facilita a identificação de padrões e anomalias específicas. Já ao nível do valor individual, cada dado é avaliado isoladamente, permitindo uma análise pormenorizada da sua validade e coerência.

As métricas são geralmente construídas para quantificar diretamente as DQD (23). Nesse sentido, Goknil et al. (9) identificaram 41 métricas distintas distribuídas por 17 dimensões, no contexto de sistemas IdC e SCF. As métricas descritas a seguir foram reunidas e analisadas por Peixoto et al. (1), com o objetivo de quantificar, de forma prática e adaptada ao contexto da I4.0, as principais DQD previamente discutidas.

2.5.1 Métricas para medir a Acurácia

A acurácia representa o grau de conformidade entre os dados armazenados e os seus valores reais, refletindo a veracidade da informação em relação ao mundo físico que pretende descrever. Esta dimensão é crítica em contextos industriais, onde decisões operacionais e estratégicas são frequentemente automatizadas e altamente dependentes de dados provenientes de sensores. Erros nesta dimensão podem derivar de medições imprecisas, interferência eletromagnética, sensores defeituosos ou falhas na transmissão de dados (1; 16).

No contexto da I4.0, Peixoto et al. (1) propuseram três métricas distintas para avaliar a acurácia: duas baseadas na abordagem de Mahanti (23), aplicáveis a diferentes níveis de granularidade, e uma terceira derivada do conjunto de métricas sistematizadas por Goknil et al. (9).

Acurácia ao nível do registo

A avaliação da acurácia ao nível do registo visa determinar se um conjunto completo de valores associados a um mesmo instante temporal representa fielmente à realidade (1). Em ambientes industriais, este tipo de análise é crucial, dado que múltiplos sensores recolhem simultaneamente diferentes parâmetros, como temperatura, pressão ou vibração, e a fiabilidade da análise depende da validade conjunta de todos os campos. Se os valores dos elementos críticos de um registo estiverem dentro de um intervalo aceitável em relação aos valores correspondentes no registo de referência, esse registo é considerado preciso.

Neste contexto, Mahanti (23) propõe uma métrica (Eq. 2.1) baseada na contagem de registos que cumprem integralmente os critérios estabelecidos. A acurácia é então expressa como a razão entre o número de registos considerados totalmente precisos ($N_{precisos}$) e o número total de registos analisados (N_{total}):

$$\text{Acurácia} = \frac{N_{precisos}}{N_{total}} \tag{2.1}$$

Para determinar $N_{precisos}$, cada valor deve ser comparado com um conjunto de dados de referência apropriado. No caso da precisão sintática, essa comparação verifica se os valores dos elementos críticos do registo estão em conformidade com os valores permitidos no domínio, como garantir que uma temperatura registada esteja dentro de um intervalo aceitável, enquanto no caso da precisão semântica, a comparação é feita com o valor real correspondente no mundo real, como verificar se a temperatura registada pelo sensor corresponde ao valor real medido.

Para que um registo seja classificado como "preciso", todos os seus elementos críticos devem estar em conformidade com os valores definidos como válidos. Esta conformidade pode ser avaliada de forma sintática, isto é, através da verificação de que os valores se encontram dentro dos limites permitidos do domínio (por exemplo, uma temperatura entre 0^{o}C e 100^{o}C), ou de forma semântica, comparando os valores registados com valores de referência externos ou medidos por outros dispositivos de confiança.

A avaliação desta métrica, de acordo com os critérios definidos por Loshin (15), é apresentada no artigo de Peixoto et al. (1). No caso da métrica de acurácia ao nível do registo, destaca-se a sua mensurabilidade, pois pode ser facilmente aplicada a qualquer conjunto de dados que disponha de uma referência válida para comparação. No entanto, a sua

capacidade de detalhe é limitada, uma vez que, apesar de identificar registos imprecisos, não indica diretamente quais os campos responsáveis, o que dificulta o diagnóstico e a correção de falhas específicas.

Embora esta métrica seja particularmente útil em sistemas onde a uniformidade dos dados é imperativa, apresenta também limitações importantes. A principal reside na sua rigidez: basta um único valor incorreto para invalidar todo o registo, o que pode penalizar excessivamente sistemas com variação natural nos dados ou suscetíveis a pequenas flutuações não críticas.

Acurácia ao nível do elemento

A métrica de acurácia ao nível do elemento centra-se na avaliação individual de cada atributo registado, como a temperatura. Em contraste com a abordagem ao nível do registo, esta métrica permite uma análise mais granular, identificando quais os elementos específicos que apresentam desvios ou erros, sem comprometer a totalidade do registo. Foca-se na avaliação individual de cada campo, útil para localizar atributos específicos problemáticos.

A acurácia dos elementos de dados é avaliada comparando cada elemento com o seu domínio de valores definido. É idêntico à equação 2.1, mas é contextualizado no elemento de dados, onde $N_{precisos}$ é o número de valores precisos do elemento avaliado e N_{total} é o número total de elementos de dados.

A avaliação completa desta métrica é apresentada em (1), onde sobressaem a sua capacidade de detalhe, que permite localizar com precisão os atributos responsáveis por falhas na qualidade dos dados, e a rastreabilidade, que possibilita acompanhar a evolução da acurácia de cada campo ao longo do tempo, sendo particularmente útil para identificar a degradação de sensores em ambientes industriais. Esta métrica é especialmente relevante em contextos industriais, onde os dados são recolhidos em larga escala por sensores e a precisão de cada elemento individual pode impactar diretamente a qualidade do controlo de processos ou a eficácia da manutenção preditiva. Para além de permitir uma análise detalhada da acurácia de cada atributo, facilita a identificação de padrões de erro recorrentes associados a sensores específicos, possibilitando intervenções corretivas mais direcionadas. A métrica definida pela equação 2.1 pode ser aplicada na avaliação da qualidade dos dados em ambientes de IoT, desde que sejam efetuadas as devidas adaptações aos domínios de referência, tendo em conta a natureza dinâmica destes sistemas (1).

No estudo conduzido por Goknil et al. (9), foram identificadas várias métricas utilizadas na I4.0, entre as quais uma equivalente à apresentada, classificada como M5 e originalmente proposta por Byabazaire et al. (51). A principal diferença reside na abordagem inversa: em vez de contabilizar os valores corretos, a métrica M5 calcula a proporção de valores incorretos e subtrai esse valor de 1, obtendo assim uma estimativa do grau de acurácia dos dados.

Acurácia ao nível do valor

A terceira métrica apresentada por Peixoto et al. (1) para a avaliação da acurácia foca-se na análise de cada valor individual, quantificando a sua proximidade relativa em relação ao intervalo de valores considerados válidos. Esta abordagem, identificada como M4 no estudo sistemático de Goknil et al. (9) e originalmente proposta por Sicari et al. (52),

revela-se particularmente eficaz na deteção de *outliers* e desvios em ambientes dinâmicos, onde os dados são atualizados de forma contínua e sujeitos a flutuações naturais.

A equação 2.2 permite normalizar cada valor x_i em relação ao conjunto X dos valores válidos disponíveis:

$$z_i = \frac{x_i - \min(X)}{\max(X) - \min(X)} \tag{2.2}$$

O resultado z_i varia, idealmente, entre 0 e 1. Valores fora deste intervalo sugerem que x_i encontra-se fora do domínio observado, podendo indicar um erro, anomalia ou leitura extrema. Esta métrica não se foca apenas em binarizar a acurácia (certo/errado), mas sim em quantificar o quão distante um valor está dos limites esperados, fornecendo, assim, uma visão mais graduada do desvio.

A sua aplicação é particularmente útil em sistemas onde a análise de tendências e a deteção de anomalias são cruciais, como em manutenção preditiva. Por exemplo, se a pressão de uma máquina tende progressivamente a aproximar-se do limite superior, o valor de z_i refletirá essa tendência antes de se atingir um estado crítico.

A avaliação desta métrica é apresentada em (1), tendo sido aplicadas as respetivas características de uma boa métrica, conforme identificado por Loshin (15). Entre os critérios avaliados, destaca-se a clareza na definição, uma vez que a métrica se baseia numa fórmula de normalização, com valores entre 0 e 1 que facilitam a interpretação de desvios. A sua relevância para o negócio é evidente em sistemas de manutenção preditiva, permitindo antecipar falhas com base em variações progressivas. Além disso, apresenta uma forte rastreabilidade, já que possibilita o acompanhamento da evolução dos dados ao longo do tempo, desde que os limites de referência sejam periodicamente ajustados para refletir as condições reais do sistema.

Em síntese, as três métricas de acurácia analisadas oferecem perspetivas complementares, adaptando-se a diferentes exigências de monitorização da qualidade dos dados. A métrica ao nível do registo (Eq. 2.1) fornece uma visão agregada da qualidade dos dados, avaliando a conformidade de cada registo com os seus valores de referência. No entanto, a sua exigência de acurácia total pode ser excessiva em contextos industriais com variabilidade natural. A métrica ao nível do elemento permite uma análise mais granular, focando-se em atributos específicos dentro de cada registo, sendo eficaz na identificação de falhas localizadas. No entanto, depende fortemente da estabilidade dos domínios de referência e da existência de padrões bem definidos. Por fim, a métrica ao nível do valor (Eq. 2.2) destaca-se pela sua adaptabilidade a ambientes dinâmicos. Avaliando a posição relativa de cada valor face aos intervalos observados, permite detetar desvios ligeiros que podem sinalizar falhas incipientes em sensores ou equipamentos, contribuindo para ações de manutenção preditiva. A sua natureza sensível a variações individuais torna-a particularmente útil para o acompanhamento contínuo da variabilidade dos dados e deteção precoce de anomalias (1).

2.5.2 Métricas para medir a Completude

A dimensão da completude refere-se à presença ou ausência de dados esperados (23), ou seja, à medida em que os valores necessários estão efetivamente disponíveis no con-

junto de dados. Esta dimensão é particularmente crítica em cenários industriais e em sistemas baseados em IdC e SCF, onde falhas em sensores, desligamentos programados ou interrupções na recolha podem originar valores em falta ou nulos, comprometendo a fiabilidade das análises subsequentes (1; 16). Mahanti (23) propõe uma distinção da completude em três níveis de avaliação: elemento de dados, registo e conjunto de dados. Contudo, considerando que, em ambientes IdC, os dados são provenientes de múltiplas fontes e dispositivos, torna-se necessário adotar uma análise mais detalhada (1). Assim, a avaliação ao nível do conjunto de dados revela-se pouco informativa para este contexto e será, por isso, excluída desta análise.

Para que a avaliação da completude seja eficaz, é necessário identificar se cada atributo é de preenchimento obrigatório ou condicional. Atributos obrigatórios devem conter sempre valores válidos, enquanto atributos contextuais ou condicionais apenas são preenchidos quando determinadas situações ocorrem (23). Adicionalmente, é importante reconhecer os diferentes formatos com que os dados ausentes podem surgir, para além de campos nulos ou em branco, valores como "NA", "N/A", "não aplicável" ou "desconhecido" devem ser tratados como equivalentes a dados em falta.

Neste sentido, Peixoto et al. (1) propõem a utilização de três métricas complementares, cada uma aplicada a um nível distinto de granularidade: elemento de dados, registo de dados e cobertura temporal de eventos. As duas primeiras têm origem no estudo de Mahanti (23), enquanto a terceira foi introduzida pelos próprios autores.

Completude ao nível do registo

Esta métrica avalia, para cada registo, se todos os campos obrigatórios estão preenchidos. É útil para identificar registos incompletos que possam comprometer decisões baseadas na sua análise. Considerando que os atributos obrigatórios foram previamente definidos e que os valores ausentes (como nulos, em branco ou representações equivalentes) foram devidamente identificados, a métrica é calculada com base na razão entre os valores presentes num registo e o total de valores esperados num registo. A equação correspondente é a seguinte:

$$\text{Completude} = \frac{N_{total} - N_{falta}}{N_{total}} \tag{2.3}$$

Onde N_{falta} representa o número de valores em falta no registo e N_{total} o número total de valores esperados no registo.

A tabela que apresenta a avaliação das características desta métrica está presente no artigo (1). Esta métrica destaca-se pela sua clareza na definição, avaliando de forma direta a presença ou ausência de campos num registo, com um resultado que varia entre 0 (completamente incompleto) e 1 (totalmente completo). É facilmente mensurável, bastando verificar se os campos obrigatórios estão preenchidos, e possui grande relevância para o negócio, sobretudo em contextos onde decisões operacionais dependem de dados críticos e completos. No entanto, embora seja útil para identificar registos incompletos, tem limitações ao nível da capacidade de detalhe, pois não indica diretamente quais os campos em falta, exigindo uma análise adicional para diagnóstico mais aprofundado.

Completude ao nível do elemento

Na completude ao nível do elemento, avalia-se a presença de valores para cada atributo específico (ex: temperatura, pressão). A equação é igual à do nível do registo (Eq. 2.3), mas a granularidade é agora ao nível de cada elemento. Onde N_{falta} representa o número de valores em falta e N_{total} o número total de valores esperados.

A métrica é analisada segundo os critérios de qualidade propostos por Loshin (15), conforme apresentado em (1). Esta métrica destaca-se pela sua mensurabilidade, uma vez que a proporção de valores presentes em relação ao total esperado é facilmente quantificável e aplicável mesmo em ambientes com elevada volumetria de dados. Outro ponto forte é a sua rastreabilidade, já que permite monitorizar a evolução da completude de cada atributo ao longo do tempo, identificando tendências de degradação em sistemas de aquisição contínua. No entanto, apresenta também uma limitação ao nível da capacidade de detalhe, pois apesar de permitir localizar elementos incompletos, não fornece uma explicação direta para a ausência dos dados, exigindo análises complementares para diagnóstico da causa.

Esta abordagem é também referida por Goknil et al. (9), que identificam métricas semelhantes designadas por M10, M11 e M13, com origem nos trabalhos de Kuemper et al. (13), Sicari et al. (52) e Byabazaire et al. (51), respetivamente. Em todas estas variações, a lógica baseia-se na contagem de valores ausentes para inferir o grau de completude.

Completude baseada na ocorrência temporal

A última métrica para medir a completude, apresentada por Peixoto et al. (1), tem como objetivo avaliar a regularidade com que os eventos ocorrem ao longo de intervalos de tempo pré-definidos, por exemplo, registos esperados a cada 5 minutos. Ao invés de se centrar em valores ausentes dentro de registos individuais, esta métrica analisa a presença ou ausência de eventos completos dentro de cada janela temporal. Trata-se de uma abordagem eficaz para identificar falhas de recolha, perdas de dados, interrupções ou mesmo duplicações no fluxo de eventos. A métrica é calculada com base na razão entre o número de ocorrências reais registadas (N_{occur}) e o número de ocorrências esperadas (N_{exp}) num determinado intervalo de tempo T, conforme apresentado na seguinte equação:

$$\text{Completude} = \frac{N_{ocorr}}{N_{exp}} \tag{2.4}$$

Nesta abordagem, é necessário definir previamente o número esperado de eventos ocorridos por intervalo de tempo, com base no comportamento previsto do sistema ou na frequência de envio dos sensores. A comparação entre o número efetivamente observado e o número esperado permite identificar intervalos com ausência total de dados ou picos anómalos de atividade.

Esta métrica, centrada na avaliação da cobertura temporal das ocorrências, deve ser analisada segundo os critérios de qualidade de métricas propostos por Loshin (15), conforme sintetizado em (1). Destaca-se pela sua relevância para o negócio, sendo particularmente eficaz em contextos industriais onde a recolha contínua de dados é crítica para garantir a fiabilidade dos processos. Também apresenta elevada mensurabilidade, já que a comparação entre o número de eventos registados e o número esperado num dado intervalo

é de fácil cálculo. Além disso, outro dos seus pontos fortes é a representação, já que os resultados podem ser facilmente visualizados através de gráficos temporais, permitindo identificar rapidamente falhas, lacunas ou padrões irregulares na recolha de dados. Contudo, apresenta limitações ao nível da capacidade de detalhe, uma vez que, embora permita localizar precisamente os intervalos com falhas, não fornece diretamente a causa do problema (ex: erro de sensor, falha de comunicação ou problema de sincronização).

A análise das diferentes métricas de completude evidencia que todas desempenham um papel complementar e indispensável na avaliação da qualidade dos dados em ambientes de I4.0 (1). Em primeiro lugar, a métrica de completude ao nível do elemento (Eq. 2.3) permite identificar atributos específicos que contêm valores em falta, fornecendo uma visão detalhada sobre a integridade de cada parâmetro individual. Esta granularidade é essencial em operações industriais, onde a ausência de dados críticos, como temperatura, pressão ou velocidade, pode comprometer a atuação em tempo real. Aplicada ao nível do registo (também Eq. 2.3, mas com enfoque no conjunto de atributos), esta mesma métrica permite avaliar se todos os campos essenciais estão presentes num dado registo, apoiando decisões importantes relacionadas com falhas operacionais ou interrupções de equipamentos. Por outro lado, a métrica baseada na ocorrência temporal (Eq. 2.4) destaca-se pela sua capacidade de monitorizar a regularidade da recolha de eventos ao longo do tempo. Ao comparar o número de ocorrências reais com o número esperado num determinado intervalo, esta métrica possibilita a deteção de falhas na recolha, interrupções sistemáticas, ou até duplicações causadas por reenvio indevido de dados. Tal capacidade é especialmente relevante em ambientes de I4.0, onde a continuidade da recolha de dados é essencial para assegurar rastreabilidade e respostas rápidas a desvios no processo.

Em conjunto, estas três métricas constituem uma abordagem abrangente para a avaliação da completude dos dados. Enquanto as métricas ao nível do elemento e do registo detetam lacunas no conteúdo dos dados, a métrica temporal permite avaliar a consistência da sua recolha ao longo do tempo. A sua aplicação integrada assegura não só a cobertura completa dos dados críticos, mas também a sua fiabilidade contínua, aspetos fundamentais para a otimização de processos, prevenção de falhas e apoio à tomada de decisão.

2.5.3 Métricas para medir a Consistência

A dimensão da consistência diz respeito à uniformidade e coerência dos dados, assegurando que os valores registados mantêm uma representação lógica e sincronizada da realidade ao longo do tempo e entre diferentes fontes. No contexto da I4.0, esta dimensão é particularmente relevante devido à integração de múltiplos sistemas, sensores e plataformas que recolhem e processam dados de forma distribuída e em tempo real (1). A presença de inconsistências pode comprometer a integridade das análises, induzir em erro as decisões e gerar conclusões contraditórias entre diferentes subsistemas.

Segundo Mahanti (23), a consistência pode ser analisada em três níveis distintos: ao nível do elemento, avaliando se os dados de atributos relacionados seguem regras semânticas pré-definidas, ao nível do registo cruzado, que verifica a conformidade de múltiplos registos entre diferentes fontes, e ao nível do conjunto de dados, analisando a correspondência entre o sistema de origem e o sistema de destino. Cada um destes níveis responde a um tipo específico de problema de consistência. Por exemplo, dois sensores da mesma máquina devem reportar valores em escalas compatíveis (elemento), as tabelas de registo devem referenciar os mesmos identificadores (registo cruzado) e os dados guardados numa base

de dados devem ser coerentes com os dados originais (conjunto de dados) (1).

Nas secções seguintes, apresentam-se as métricas propostas para avaliar a consistência em cada um destes níveis, conforme descrito por Peixoto et al. (1) e com base nas abordagens discutidas por Mahanti (23).

Consistência ao nível do elemento

A consistência ao nível do elemento de dados avalia se os valores de diferentes atributos dentro de um mesmo registo obedecem a regras semânticas definidas, garantindo que a informação se mantém lógica e coerente internamente (1). Esta abordagem é especialmente relevante em cenários industriais, onde múltiplos sensores monitorizam simultaneamente variáveis interdependentes (por exemplo, temperatura e pressão de um processo), sendo fundamental assegurar que os valores registados respeitam relações consistentes entre si.

Para aplicar esta métrica, é necessário identificar previamente um conjunto de regras semânticas entre os atributos, como, por exemplo, "se o modo da máquina for 'desligado', a velocidade deve ser igual a zero" ou "se a velocidade de rotação aumenta, a temperatura também deve aumentar". A consistência é então determinada pela razão entre o número de registos que cumprem as regras de consistência definidas entre elementos ($N_{validos}$) e o número total de regras existentes (N_{total}):

$$\text{Consistência} = \frac{N_{validos}}{N_{total}} \tag{2.5}$$

Peixoto et al. (1) apresentam esta métrica como uma forma eficaz de detetar erros semânticos que não são imediatamente identificáveis por métricas convencionais, uma vez que se baseiam em regras que refletem o conhecimento de domínio. A sua eficácia depende diretamente da qualidade e abrangência das regras definidas. Em sistemas complexos, a definição de regras pode tornar-se desafiante, exigindo a colaboração de especialistas no processo industrial. No entanto, quando bem implementada, esta métrica fornece uma visão clara sobre a integridade semântica dos dados.

A avaliação completa desta métrica segundo os critérios propostos por Loshin (15) encontra-se detalhada em (1). Esta métrica destaca-se pela sua relevância para o negócio, pois, em contextos industriais, a violação de relações lógicas entre atributos pode sinalizar estados de falha iminente ou comportamentos anómalos críticos. Apresenta também uma elevada reportabilidade, permitindo integrar com facilidade os resultados em relatórios técnicos e *dashboards*. No entanto, a sua mensurabilidade depende fortemente da definição clara e completa das regras semânticas entre os atributos, o que pode representar um desafio em sistemas heterogéneos ou pouco documentados.

No contexto da I4.0, onde vários sensores e sistemas são integrados, esta métrica continua a ser uma excelente escolha, pois oferece uma visão detalhada e flexível, permitindo que inconsistências sejam identificadas em relações específicas entre elementos de dados, algo crucial para manter a consistência e o desempenho dos sistemas em tempo real.

Consistência ao nível de registos cruzados

A consistência ao nível de registos cruzados (também designada por consistência entre registos) visa assegurar que múltiplos registos, provenientes de diferentes fontes ou siste-

mas, mantêm coerência entre si com base em regras de integridade semântica previamente estabelecidas (1; 23). Esta verificação é particularmente relevante em ambientes industriais, onde dados inter-relacionados são frequentemente gerados e armazenados por diversos sistemas, como sensores, bases de dados operacionais ou plataformas de supervisão. Nesses contextos, incoerências entre registos podem comprometer a rastreabilidade e originar decisões incorretas.

A métrica correspondente baseia-se na aplicação coletiva de um conjunto de regras semânticas que definem as relações esperadas entre os dados distribuídos. Após a identificação dessas regras, identifica-se o número de registos que cumprem integralmente os critérios definidos ($N_{validos}$) e o número total de registos analisados (N_{total}). A consistência global entre registos é então quantificada através da equação 2.5. Esta abordagem, apresentada por Peixoto et al. (1), está em conformidade com a estrutura metodológica proposta por Mahanti (23).

A avaliação desta métrica, de acordo com os critérios definidos por Loshin (15), é apresentada no artigo de Peixoto et al. (1). A principal utilidade reside na capacidade de detetar incoerências que resultam da fusão de dados heterogéneos, como, por exemplo, diferenças em identificadores de equipamentos, ou valores contraditórios entre sistemas. Embora seja eficaz para validar a integridade entre conjuntos de registos, a métrica depende fortemente da definição clara de regras semânticas e da existência de relacionamentos explícitos entre os dados.

Consistência ao nível do conjunto de dados

A consistência ao nível do dataset diz respeito à conformidade entre os dados presentes no sistema de origem e aqueles efetivamente armazenados no sistema de destino, como consequência de processos de transferência, transformação ou carregamento de dados (23). Em ambientes industriais, onde a recolha e integração de dados ocorre em tempo real e de forma distribuída, falhas de sincronização, erros nos processos de carga ou reprocessamentos incompletos podem originar discrepâncias substanciais entre o conteúdo original e o final (1). Tais inconsistências surgem frequentemente quando apenas uma parte dos dados é carregada ou quando o recarregamento não é efetuado a partir do ponto de verificação mais recente, comprometendo a integridade do sistema de destino face à origem (23).

Para quantificar esta discrepância, Peixoto et al. (1), com base na métrica proposta por Mahanti (23), apresentam o cálculo da inconsistência através do racio entre a diferença entre o número de registos na origem (N_{origem}) e no destino ($N_{destino}$), e o total de registos esperados na origem (N_{origem}). A fórmula expressa a inconsistência relativa, como segue:

$$\text{Inconsistência} = \frac{|N_{\text{origem}} - N_{\text{destino}}|}{N_{\text{origem}}} \tag{2.6}$$

A origem refere-se ao sistema responsável pela geração e transmissão dos dados, que pode incluir o próprio sensor ou um sistema intermediário que agrega os dados antes da transmissão, o sistema de destino é o local de armazenamento final. Neste caso, o valor ideal da métrica é zero, o que indicaria correspondência exata entre as duas fontes. Valores positivos indicam perda ou duplicação de dados, sendo essencial investigar a origem da

discrepância: se por falha de comunicação, erro no processo de ingestão, ou problemas com checkpoints e atualizações parciais (23).

A abordagem de avaliação segue os princípios de Loshin (15), posteriormente aplicados ao contexto da indústria 4.0 por Peixoto et al. (1). Esta métrica destaca-se pela sua relevância para o negócio, sobretudo em contextos da I4.0, onde a consistência na transferência de dados entre sistemas é essencial para garantir a integridade da informação ao longo de *pipelines* distribuídos. Apresenta também uma boa rastreabilidade, permitindo acompanhar ao longo do tempo a evolução da consistência entre sistemas e identificar padrões de falhas ou melhorias. No entanto, a sua mensurabilidade pode ser desafiante em alguns cenários, especialmente quando os sistemas de origem não disponibilizam mecanismos diretos e fiáveis de contagem de registos, o que compromete a precisão da comparação com os dados de destino.

Dados os desafios impostos pela I4.0, onde a consistência da informação é essencial para decisões em tempo real, a dimensão da consistência assume um papel central na garantia da qualidade dos dados. As três métricas analisadas, ao nível do elemento, entre registos e ao nível do dataset, fornecem abordagens complementares que abrangem desde a coerência interna até à integridade global dos dados ao longo de todo o sistema (1). A métrica de consistência ao nível dos elementos de dados (Eq. 2.5) permite verificar a lógica interna de cada registo, garantindo que atributos relacionados mantêm coerência semântica. Esta análise é fundamental para detetar erros de configuração, medições incompatíveis ou anomalias lógicas em registos provenientes de sensores. Por sua vez, a métrica de consistência entre registos (mesma Eq. 2.5, aplicada entre conjuntos de dados distintos) avalia a conformidade de múltiplos registos provenientes de fontes diversas, sendo essencial em ambientes onde a informação é gerada por diferentes sistemas, mas precisa de se manter sincronizada. Já a métrica ao nível do dataset (Eq. 2.6) assegura a integridade durante processos de extração, transformação e carregamento, detetando perdas, duplicações ou falhas na sincronização entre sistemas de origem e destino.

Estas três métricas fornecem uma abordagem abrangente para garantir a consistência dos dados em todo o ciclo de vida da informação. Permitem detetar incoerências locais e globais, apoiar processos de validação automática e reduzir o risco de decisões baseadas em dados inconsistentes. No entanto, a aplicabilidade e a utilidade de cada métrica dependem sempre do contexto específico em que são implementadas, nomeadamente do tipo de sistema, da frequência de atualização dos dados e do grau de interdependência entre as fontes de informação (1).

2.5.4 Métricas para medir a Atualidade

A dimensão da atualidade refere-se ao grau em que os dados estão atualizados em função do momento em que são utilizados e aborda questões relacionadas com dados desatualizados ou obsoletos, bem como o desafio do alinhamento temporal (16). Esta dimensão é crítica em contextos industriais e de IdC (23), onde decisões operacionais dependem da disponibilidade de dados em tempo real (1). A latência entre a ocorrência de um evento no mundo físico e a disponibilização da informação correspondente pode comprometer a eficácia de sistemas de controlo ou manutenção preditiva.

À semelhança da abordagem de Peixoto et al. (1), a dimensão da atualidade será explorada através de duas métricas distintas: uma baseada na diferença temporal entre a

leitura do sensor e a disponibilização dos dados (23), e outra que integra os conceitos de *currency* e volatilidade (9; 53) para produzir um indicador composto de atualidade.

Atualidade com base no intervalo de tempo

A primeira métrica da atualidade apresentada por Mahanti (23) mede o atraso total entre o momento em que um evento ocorre e o instante em que a informação correspondente é efetivamente disponibilizada para uso. Esta abordagem é especialmente útil em sistemas industriais que exigem respostas rápidas baseadas em dados em tempo real ou quase real.

Para esta métrica, consideram-se três componentes temporais: o momento em que o evento ocorre no mundo físico ($D_{\text{ocorrência}}$), o momento em que os dados são processados ou disponibilizados no sistema ($D_{disponibilizado}$) e o momento em que os dados são efetivamente utilizados ou acedidos pelo utilizador/sistema ($D_{entrega}$). A equação utilizada por Mahanti (23) para calcular a atualidade é:

$$\text{Atualidade} = (D_{\text{entrega}} - D_{\text{disponibilizado}}) + (D_{\text{disponibilizado}} - D_{\text{ocorrência}}) \tag{2.7}$$

Em muitos casos, assume-se que $D_{entrega} = D_{disponibilizado}$, simplificando o cálculo para o atraso entre a ocorrência do evento e a sua disponibilização. Contudo, quando existe uma diferença significativa entre os momentos de disponibilização e utilização, a métrica permite uma avaliação mais rigorosa. Assim, é importante considerar o intervalo de tempo entre cada um dos componentes temporais. Esta análise detalhada permite compreender melhor os atrasos e o impacto de vários fatores na relevância dos dados, proporcionando uma visão mais granular da atualidade global.

Esta métrica requer ainda a definição de um limiar máximo aceitável de atraso, que depende do contexto específico de aplicação, uma vez que, mesmo dentro da mesma fábrica, máquinas diferentes podem ter tempos de operação distintos. Essas variações nos ciclos operacionais implicam que o intervalo para considerar os dados como atuais pode ser diferente, resultando na necessidade de ajustes nos critérios de validade de acordo com os vários sistemas e equipamentos.

Esta métrica foi analisada segundo os critérios propostos por Loshin (15), apresentado por Peixoto et al. (1), e destaca-se pela sua relevância para o negócio, uma vez que, em sistemas industriais e analíticos, a atualidade dos dados é essencial para garantir decisões informadas e atempadas. Apresenta também boa representação, sendo facilmente visualizável. No entanto, a sua mensurabilidade pode ser comprometida em ambientes onde os sistemas não registam corretamente os carimbos temporais em todas as etapas do fluxo de dados, o que dificulta uma avaliação precisa da atualidade.

Atualidade com base na *Currency* e Volatilidade

A segunda métrica da atualidade é decomposta por três dimensões: *currency*, volatilidade e atualidade (24). A *currency* refere-se ao tempo decorrido desde a última atualização de um dado, refletindo a rapidez com que este é renovado após alterações no mundo real. Quando os dados são atualizados com frequência constante, a sua medição é simples, no entanto, em cenários onde essa frequência é variável, pode ser necessário estimar uma média para representar esse comportamento (1; 23). A volatilidade, por sua vez, expressa o intervalo de tempo durante o qual um dado permanece válido ou útil para análise.

Este intervalo pode variar significativamente consoante o tipo de dados e o contexto de aplicação. Em ambientes industriais com decisões em tempo real, como na I4.0, a volatilidade tende a ser reduzida (1; 23). Por último, a atualidade é entendida como a capacidade dos dados estarem não apenas atualizados, mas também disponíveis no momento certo para apoiar as ações ou eventos que motivam a sua utilização (23). Neste sentido, Ballou et al. (53) propõem uma métrica composta que considera simultaneamente a *currency* e a volatilidade como determinantes do grau de atualidade dos dados.

A métrica é composta por duas fórmulas. Primeiro, calcula-se a *currency* dos dados:

$$Currency = Idade + (Tempo_{entrega} - Tempo_{recolha}) \tag{2.8}$$

Em seguida, a atualidade é calculada como:

$$Atualidade = \max(0, (1 - Currency/Volatilidade)) \tag{2.9}$$

Nesta fórmula, a volatilidade representa o período durante o qual os dados continuam válidos e úteis. Quanto mais recente for a informação (ou maior for a sua validade temporal), mais próximo de 1 será o valor da métrica. Se a atualidade dos dados ultrapassar a sua volatilidade, o valor da atualidade será 0, refletindo que os dados estão obsoletos.

Esta métrica oferece uma visão dinâmica da qualidade temporal dos dados, adaptando-se à frequência de atualização de diferentes fontes e ao seu grau de sensibilidade temporal (1). É especialmente indicada para sistemas de produção complexos, onde diferentes sensores ou variáveis têm ciclos de atualização distintos.

Goknil et al. (9) identificaram uma métrica semelhante, designada como M28, originalmente utilizada por Sicari et al. (52). A principal distinção entre esta métrica e a expressa na equação 2.9 reside no numerador da fração utilizada para calcular a atualidade. Na abordagem de Sicari et al., a *currency* é definida diretamente como a idade dos dados, isto é, o intervalo de tempo entre a recolha da informação e o momento da sua análise. Já na fórmula composta proposta por Ballou et al. (53), essa idade é incorporada explicitamente na equação como um componente da *currency*, tornando a definição mais formal e abrangente. Apesar dessa diferença estrutural, ambas as métricas produzem o mesmo resultado, uma vez que a definição de idade adotada por Sicari et al. equivale, na prática, à *currency* formalizada por Ballou et al. (53).

Com base nos critérios de Loshin (15), a avaliação desta métrica foi desenvolvida por Peixoto et al. (1). Esta métrica destaca-se pela sua controlabilidade, permitindo ajustar a frequência de atualização dos dados ou modificar as decisões automatizadas com base na validade temporal da informação, o que é fundamental em sistemas sensíveis ao tempo. Apresenta também uma boa capacidade de detalhe, já que possibilita identificar causas associadas à desatualização, como atrasos na aquisição ou falhas na entrega dos dados. No entanto, a sua mensurabilidade pode ser limitada em cenários onde não se dispõe de definições claras de volatilidade para cada tipo de dado, ou em sistemas que não mantêm registos temporais consistentes.

As duas métricas analisadas para a avaliação da atualidade, a baseada em intervalos de tempo (Eq. 2.7) e a métrica composta que integra *currency* e volatilidade (Eq. 2.9), representam abordagens distintas, cada uma com utilidade específica conforme o contexto

de aplicação (1). A primeira, de natureza mais simples, mede diretamente o tempo decorrido entre a geração dos dados e o momento em que são utilizados. Esta métrica é eficaz na monitorização de atrasos em sistemas de transmissão em tempo real, sendo de fácil implementação e interpretação, mesmo por operadores sem formação técnica aprofundada. Em contrapartida, a métrica composta fornece uma análise mais abrangente da qualidade temporal dos dados. Ao integrar a idade dos dados com a sua janela de validade, permite não apenas verificar se os dados são recentes, mas também se continuam válidos no momento da sua utilização. Esta abordagem gera um valor normalizado entre 0 e 1, facilitando a sua representação visual e a comparação entre diferentes sensores ou sistemas com cadências de atualização distintas.

Importa destacar que estas métricas não são equivalentes. A métrica baseada em intervalo fornece uma medição direta e pontual do atraso, enquanto a métrica composta oferece uma perspetiva sofisticada e adaptável, adequada à realidade dinâmica dos SCF (1).

Para que estas e outras métricas possam ser corretamente calculadas, é necessário dispor de informação detalhada sobre os próprios dados. Esta informação pode ser extraída através de processos de *data profiling*, que desempenha um papel fundamental na avaliação da qualidade dos dados.

2.6 *Data Profiling*

Neste capítulo, analisa-se o papel do *data profiling* como elemento de suporte essencial ao cálculo das métricas da qualidade dos dados. Através da análise exploratória e sistemática dos dados, esta prática permite recolher metainformação estruturada que constitui a base para quantificar dimensões (1). A sua aplicação é particularmente relevante em ambientes industriais, onde os dados são gerados de forma contínua e heterogénea, exigindo mecanismos automatizados para garantir fiabilidade e robustez nos processos de monitorização (3).

Na I4.0, os métodos de produção tradicionais são progressivamente substituídos por operações interconectadas e adaptativas, baseadas na integração de sensores, dispositivos inteligentes e redes digitais (54). Esta interligação permite a monitorização e otimização contínua dos processos, reduzindo tempos de paragem e desperdício, e aumentando significativamente a eficiência global (55). No entanto, o sucesso desta automatização depende da qualidade dos dados, frequentemente ameaçada por leituras incorretas, ruído, falhas de transmissão ou inconsistências nos dados recolhidos (22). O *data profiling* surge, neste contexto, como uma ferramenta fundamental para a verificação sistemática da integridade dos dados, permitindo diagnosticar problemas em tempo útil e garantir a operacionalidade dos sistemas industriais (56). A sua aplicação em tempo real reduz a necessidade de intervenção manual, assegurando que os fluxos de dados mantêm níveis adequados de fiabilidade para suportar decisões estratégicas (56).

O *data profiling* assume um papel central na garantia da fiabilidade da informação utilizada em processos de tomada de decisão (19). Trata-se de uma abordagem metodológica que visa a análise detalhada dos dados com o intuito de extrair metainformação e identificar características intrínsecas de cada conjunto (18). É amplamente reconhecido como uma etapa essencial que precede qualquer atividade relacionada com a exploração de dados, como a integração, análise, monitorização da qualidade ou apoio à decisão (57). As operações típicas de *data profiling* incluem a contagem de valores distintos ou ausen-

tes, a validação de tipos de dados por coluna e a identificação de padrões recorrentes e frequências (58). Estas análises permitem revelar a estrutura subjacente dos dados e avaliar a sua consistência, facilitando a deteção de problemas como valores inválidos, duplicados ou fora dos domínios esperados. Além disso, os metadados obtidos constituem a base para o cálculo de diversas métricas de qualidade dos dados (3): a contagem de valores nulos, por exemplo, suporta a avaliação da completude, e a identificação dos tipos de dados associados a cada campo é essencial para a aplicação correta das métricas em todas as dimensões. Deste modo, o *data profiling* não só permite identificar falhas estruturais, como também contribui diretamente para a operacionalização das métricas de qualidade (3).

Segundo Abedjan et al. (18), o *data profiling* deve ser realizado tanto antes como depois das transformações nos dados, embora o momento exato da sua aplicação dependa fortemente do contexto e dos objetivos específicos da análise. Os autores explicam que, quando o objetivo é compreender a estrutura e qualidade de dados recém-obtidos, o *data profiling* deve ocorrer antes de qualquer processamento, apoiando atividades como a limpeza, integração ou otimização de consultas. Após cada transformação ou etapa de limpeza, uma nova análise torna-se essencial para validar os efeitos produzidos e revelar padrões ou dependências anteriormente ocultos. Em cenários onde os dados são altamente dinâmicos, como fluxos em tempo real ou sistemas transacionais, o *data profiling* pode assumir formas incrementais ou contínuas, sendo atualizado em paralelo com a geração e modificação dos dados. Assim, embora seja mais eficaz quando adotado como prática recorrente ao longo do ciclo de vida dos dados, o seu momento e frequência devem ser ajustados conforme os requisitos e finalidades de cada aplicação. Esta abordagem torna-se especialmente relevante em ambientes industriais com elevada frequência de aquisição, assegurando uma monitorização eficaz da qualidade dos dados em *pipelines* operacionais (1).

Apesar da sua importância, o *data profiling* enfrenta três desafios principais: ingestão, computação e interpretação dos resultados (18). O primeiro diz respeito à preparação eficiente dos dados provenientes de múltiplas fontes. O segundo prende-se com a complexidade algorítmica do processo, que tende a escalar com o número de linhas e colunas, podendo atingir complexidade exponencial quando envolvem combinações de atributos (11). Finalmente, o terceiro desafio, e possivelmente o mais crítico, é a interpretação dos resultados. Os metadados descobertos podem ser aplicados para melhorar a qualidade dos dados, traduzindo padrões e dependências em restrições ou regras para validação, limpeza e integração.

Adicionalmente, o *data profiling* desempenha um papel fundamental em diversas áreas estratégicas. Contribui para a avaliação da qualidade dos dados (57), garante a conformidade com normas de proteção de dados (59), apoia o ciclo de vida dos dados e a definição de políticas de governança (60), e permite reduzir erros, retrabalho e custos operacionais em projetos analíticos (61). De forma já antecipada, Kimball (62) defendia que o *data profiling* deveria ser uma prática obrigatória em qualquer projeto de *data warehouse*, logo após a recolha dos requisitos de negócio.

O *data profiling* não corresponde a uma tarefa única e isolada, mas sim a um conjunto diversificado de operações que podem ser agrupadas em três categorias principais: análise de colunas individuais, descoberta de dependências entre colunas e análise de dados não relacionais, como estruturas em árvore, grafos ou texto (18). A abordagem mais comum incide sobre a análise de colunas individuais, onde se encontra a categoria da cardinali-

dade, responsável por medir estatísticas fundamentais como o número total de registos, a contagem de valores nulos, os valores distintos, entre outros aspetos relevantes (18). Para além destas tarefas básicas, o *data profiling* contempla ainda a análise da distribuição de valores, através de histogramas, identificação de extremos, quartis, constância e até da análise do primeiro dígito, bem como a verificação de tipos de dados, padrões e domínios, que abrange a identificação do tipo declarado, o comprimento dos valores, o número de casas decimais, padrões regulares e a classificação semântica, como nomes próprios, códigos ou localizações.

Estes metadados são cruciais para revelar a estrutura dos dados e detetar problemas como duplicação, ausência de valores ou incoerência de formatos. Além disso, estas operações complementares permitem inferir domínios válidos, validar regras de negócio e sugerir colunas candidatas a chaves primárias, tornando o *data profiling* uma ferramenta essencial para garantir a integridade e a utilidade dos dados. Estas informações alimentam diretamente mecanismos de monitorização contínua, permitindo avaliar, em tempo real, a conformidade dos dados com os critérios de qualidade definidos para cada dimensão (1; 3).

Outro conjunto crítico de tarefas de *data profiling* diz respeito à descoberta de dependências entre colunas, como dependências funcionais, dependências condicionais e chaves compostas. Estas estruturas são fundamentais para assegurar a integridade dos dados e identificar regras de negócio implícitas (18). A descoberta de combinações de colunas únicas é particularmente relevante para identificar chaves primárias em conjuntos de dados desconhecidos. Apesar da sua relevância, a descoberta automática de dependências enfrenta três grandes desafios: o número elevado de combinações possíveis, a necessidade de validação em todo o conjunto de dados, e a dificuldade em distinguir dependências reais de padrões (18).

Por fim, em resposta às exigências de escalabilidade da I4.0, técnicas de *data profiling* incremental têm ganho destaque. Estas abordagens baseiam-se na monitorização de alterações nos dados para reutilizar resultados anteriores e aplicar o processamento apenas às partes atualizadas dos conjuntos, aumentando a eficiência sem comprometer a precisão das métricas de qualidade dos dados (63).

Em síntese, o *data profiling* constitui uma etapa indispensável no ciclo de vida da qualidade dos dados, fornecendo a metainformação necessária para a definição de métricas e a deteção de problemas estruturais. O seu papel não se limita a uma fase única, pode ser aplicado antes das transformações, para identificar problemas iniciais, após as intervenções, para validar os seus efeitos, ou de forma contínua, em sistemas em tempo real. A sua correta aplicação é, por isso, um pré-requisito para a implementação eficaz de estratégias de monitorização e melhoria da qualidade dos dados, especialmente em ambientes industriais e de elevada complexidade.

2.7 Sistemas e Arquiteturas para Qualidade dos Dados

A crescente complexidade dos sistemas industriais associados à I4.0 tem impulsionado o desenvolvimento de soluções dedicadas à avaliação e monitorização da qualidade dos dados. Estes sistemas procuram assegurar que os dados recolhidos por sensores e dis-

positivos conectados sejam fiáveis, úteis e adequados ao suporte à decisão. A literatura evidencia uma diversidade de abordagens que variam em termos de arquitetura, funcionalidades e mecanismos de avaliação, refletindo a heterogeneidade dos cenários industriais e os diferentes requisitos de cada domínio de aplicação.

Segundo Goknil et al. (9), é possível classificar os sistemas de suporte à qualidade dos dados em diferentes categorias, incluindo mecanismos de validação automática, *pipelines* com deteção de anomalias, plataformas de monitorização visual e arquiteturas extensíveis com mecanismos formais de controlo. Destacam-se, por exemplo, os sistemas baseados em regras lógicas, onde os dados são avaliados com base em critérios definidos manualmente. Estes sistemas oferecem elevada interpretabilidade e simplicidade de implementação, mas tendem a ser menos adaptáveis a contextos dinâmicos e sujeitos a alterações frequentes nos padrões de funcionamento. Em contraste, abordagens baseadas em *machine learning*, supervisionado ou não supervisionado, têm vindo a ganhar importância na deteção de padrões não comuns de má qualidade, como desvios estatísticos, *outliers* ou correlações inesperadas (9). Estas técnicas apresentam maior flexibilidade e capacidade preditiva, permitindo adaptar-se a contextos variáveis. No entanto, exigem fases de treino rigorosas e estão fortemente dependentes da qualidade e representatividade dos dados históricos utilizados, o que pode limitar a sua eficácia em ambientes com variabilidade elevada ou dados incompletos (9).

No domínio da ingestão de dados industriais, Ji et al. (64) propuseram um modelo abrangente adaptado à ingestão de dados de dispositivos heterogéneos, concebido para lidar com desafios típicos da I4.0, como a diversidade de fontes, tipos de dados e requisitos de sincronização. Implementado na plataforma *Industrial Big Data Platform*, o modelo assenta na utilização de *templates* de dispositivos, que definem a estrutura de sensores e parâmetros esperados para cada tipo de dispositivo, e na aplicação coordenada de quatro estratégias complementares: sincronização de dados, segmentação temporal, divisão por parâmetros e indexação hierárquica. Estas estratégias permitem uma ingestão estruturada e escalável de dados provenientes de múltiplas origens. Em conjunto, estas estratégias conferem ao modelo de Ji et al. (64) uma solução eficaz para ingestão, armazenamento e análise de dados industriais, com validação prática em cenários reais. Contudo, importa salientar que o sistema não foi concebido para suportar aplicações em tempo real, o que limita a sua aplicabilidade em contextos com requisitos de latência reduzida e resposta imediata.

No plano mais genérico, Sawant e Shah (65) apresentaram um levantamento estruturado de padrões arquiteturais de ingestão e *streaming*, com ênfase em dados não estruturados provenientes de múltiplas fontes. Os autores identificam cinco padrões principais: extrator de múltiplas fontes, conversor de protocolo, padrão de destinos múltiplos, transformação no momento certo e padrão de transmissão em tempo real. Estes padrões são amplamente utilizados como blocos de construção para arquiteturas escaláveis de *Big Data*, aplicáveis em ambientes empresariais com elevado volume de dados semi-estruturados ou não estruturados.

Entre as ferramentas que operacionalizam esses padrões destaca-se o Gobblin (66), uma *framework* desenvolvida para unificar diferentes formas de ingestão de dados em larga escala. Com suporte para ingestão em lote e *streaming*, o Gobblin é altamente configurável e extensível para diferentes tipos de fontes e destinos. No entanto, a sua aplicação continua orientada a contextos genéricos de *Big Data*, sem integração nativa com mecanismos

específicos de avaliação de qualidade dos dados em ambientes industriais.

Complementarmente, Irfan e George (67) realizaram uma revisão sistemática sobre os desafios e ferramentas associadas à ingestão de dados em ambientes de *Big Data*. A análise abrange aspetos técnicos como latência, escalabilidade, heterogeneidade de fontes, e integração com sistemas analíticos, além de identificar diversas ferramentas e plataformas utilizadas em *pipelines* modernos. Apesar da abrangência da análise, os autores não contemplam requisitos específicos da I4.0, como a fiabilidade da ingestão em tempo real ou a monitorização contínua da qualidade dos dados.

De forma semelhante, Vyas et al. (68) propuseram um sistema melhorado para ingestão de dados no ecossistema *Hadoop*[1]. A solução apresentada supera limitações de ingestão manual, suporte a dados não estruturados e incapacidade de resposta em tempo real, típicas de arquiteturas baseadas apenas em HDFS. Ainda assim, a proposta permanece generalista, não incorporando mecanismos dedicados à avaliação da qualidade dos dados nem abordando os requisitos específicos da I4.0. Ainda assim, o trabalho reforça a importância de arquiteturas modulares e da utilização de ferramentas especializadas para suportar a crescente heterogeneidade e volume de dados nos *pipelines* modernos.

Em contraste, algumas arquiteturas recentes focadas na I4.0 integram múltiplas DQD em simultâneo, permitindo a visualização em tempo real do estado dos dados através de *dashboards* interativos (8). Estas arquiteturas combinam módulos de avaliação automática, como os disponibilizados por *frameworks* tipo *Great Expectations*, com ferramentas de visualização como Grafana[2], otimizando a capacidade de resposta dos operadores em ambiente industrial contínuo. Mecanismos adicionais, como a utilização de *t-digest* para deteção de *outliers*, e a atribuição de pontuações por dimensão de qualidade, reforçam o suporte a estratégias de manutenção preditiva e resposta a falhas em tempo útil (8).

Entre as contribuições mais relevantes neste domínio destaca-se o trabalho de Oliveira e Oliveira (69), que propõem uma arquitetura modular baseada em *Apache Kafka*[3], suportada por *plugins* de avaliação da qualidade dos dados. Esta abordagem assegura escalabilidade e flexibilidade, promovendo a separação entre ingestão, validação e armazenamento. A investigação foi posteriormente expandida por Oliveira et al. (8), que introduzem uma arquitetura completa de ingestão e monitorização contínua, integrando validação por esquemas, regras formais de qualidade com *Great Expectations*, armazenamento distribuído (MongoDB[4], Cassandra[5], InfluxDB[6]) e visualização em tempo real via Grafana. Esta solução representa um avanço significativo ao responder diretamente às necessidades da I4.0, conciliando ingestão de alto desempenho com avaliação rigorosa da qualidade dos dados em tempo real.

Concluindo, a literatura evidencia uma evolução significativa no desenvolvimento de sistemas de ingestão e monitorização de dados, refletindo o crescente impacto da I4.0 nas exigências de qualidade dos dados. Desde abordagens genéricas baseadas em *Big Data* até arquiteturas especializadas em ambientes industriais, observa-se uma tendência clara de transição de modelos centrados em processamento em lote para soluções adaptadas à

[1] `https://hadoop.apache.org/`
[2] `https://grafana.com/`
[3] `https://kafka.apache.org/`
[4] `https://www.mongodb.com/`
[5] `https://cassandra.apache.org/`
[6] `https://www.influxdata.com/`

ingestão contínua, validação automática e análise em tempo real. No entanto, muitos dos sistemas analisados continuam a carecer de mecanismos específicos para avaliação formal da qualidade dos dados, integração com métricas de qualidade ou capacidade de reação imediata a falhas.

Capítulo 3

Análise de Casos de Estudo

Este capítulo apresenta três casos de estudo desenvolvidos com o objetivo de validar, aplicar e evoluir metodologias para avaliação e monitorização da qualidade dos dados em ambientes industriais. Estes estudos constituem o pilar da componente experimental da dissertação, refletindo a crescente complexidade dos cenários abordados, o aprimoramento das soluções propostas e a sua adaptação a diferentes contextos operacionais.

A investigação teve início com a utilização de dados simulados e baseados na literatura, o que permitiu uma primeira implementação das métricas e técnicas em ambientes controlados, com um grau de incerteza reduzido. Em seguida, recorreu-se a dados públicos disponíveis online, o que permitiu testar a generalização das abordagens propostas. Finalmente, as soluções foram aplicadas a um sistema de produção real na empresa JPM[1], permitindo validar o *pipeline* completo num cenário industrial exigente, com dados heterogéneos e variabilidade operacional significativa.

Os três casos foram organizados de forma progressiva e complementar, permitindo observar a evolução das estratégias aplicadas e as melhorias incrementais introduzidas em cada fase:

- O **Caso de Estudo 1** incide sobre um processo de extrusão de plástico, com ênfase na aplicação prática das quatro DQD: acurácia, completude, consistência e atualidade. As métricas são implementadas e avaliadas em tempo real, permitindo analisar a sua adaptabilidade ao contexto industrial. Além disso, é realizada uma comparação entre diferentes métodos de deteção de anomalias, explorando a complementaridade entre abordagens baseadas em regras e técnicas estatísticas e de aprendizagem.
- O **Caso de Estudo 2**, aplicado a um sistema de bombeamento de água, introduz uma abordagem mais avançada com a construção de índices da qualidade. Estes índices são calculados com base em duas DQD, permitindo representar a fiabilidade dos dados ao longo do tempo de forma mais intuitiva e global. Esta solução facilita a monitorização contínua e a comparação entre diferentes sensores e condições operacionais.
- O **Caso de Estudo 3** aplica os conceitos desenvolvidos a um sistema industrial real da empresa JPM, mais complexo e representativo de um ambiente produtivo

[1] JPM Industry, `https://jpm.pt/en/homepage/`

típico. Este estudo permite testar a escalabilidade do *pipeline* de avaliação, bem como a eficácia das métricas anteriormente aplicadas. A integração com fluxos de dados reais permite observar o comportamento do sistema sob condições variáveis e validar a aplicabilidade prática da solução desenvolvida.

Cada estudo segue uma estrutura comum, incluindo a caracterização do cenário, a descrição do dataset, a aplicação das métricas e técnicas, as alterações ou melhorias introduzidas em relação ao estudo anterior, os resultados obtidos e a respetiva análise crítica. Esta abordagem incremental permite não só validar a aplicabilidade das metodologias propostas, como também evidenciar a sua evolução técnica e científica ao longo do desenvolvimento da dissertação. A aplicação final num caso de estudo real reforça o potencial de integração da solução em ambientes produtivos e destaca o seu contributo para a melhoria da fiabilidade dos sistemas industriais orientados por dados.

Importa salientar que o conjunto de dados utilizado no **Caso de Estudo 1** foi criado especificamente para este projeto, o que garantiu um ambiente controlado e, portanto, ideal para experimentação e afinação de métricas. Por esse motivo, este cenário foi também utilizado para comparar diferentes técnicas de deteção de anomalias, permitindo refinar as métricas inicialmente aplicadas. Dessa análise resultou uma nova abordagem para a aplicação da métrica de acurácia, mais adequada à deteção de anomalias em séries temporais. No entanto, esta evolução não se refletiu no **Caso de Estudo 2**, uma vez que o trabalho correspondente já se encontrava finalizado e aceite para publicação. Esta situação ilustra a natureza iterativa da investigação, na qual o conhecimento adquirido em fases posteriores contribuiu para o aperfeiçoamento contínuo das soluções desenvolvidas.

3.1 Caso de Estudo 1 – Validação de Métricas de Qualidade em Tempo Real

Este primeiro caso de estudo tem como objetivo demonstrar a aplicação prática de um conjunto de métricas de qualidade dos dados num ambiente industrial, recorrendo a um cenário representativo do setor de extrusão de plásticos e é baseado no artigo ***Real-Time Manufacturing Data Quality: Leveraging Data Profiling and Quality Metrics*** de Peixoto et al. (3).

Neste contexto, são analisadas quatro DQD, acurácia, completude, consistência e atualidade, com recurso a algumas métricas previamente identificadas. Os dados foram gerados de forma simulada, com base em parâmetros reais descritos em estudos prévios, refletindo um funcionamento contínuo ao longo de três dias.

A extrusão é um processo central na produção de diversos produtos plásticos, como tubos, revestimentos, isolamentos de fios e cabos, e monofilamentos (70). Neste estudo, o foco incide sobre a extrusão de parafuso único, processo no qual grânulos de plástico são transformados num fluido viscoso fundido, originando um produto final sólido ou flexível (3). Este processo envolve um parafuso rotativo com lâminas helicoidais inserido num cilindro aquecido (2). A extrusora é alimentada a partir de um funil superior, que conduz o material ao longo do cilindro. A rotação do parafuso transporta, aquece e comprime o material, que é então forçado a sair através de uma matriz que lhe dá a forma final desejada (70).

A Figura 3.1 ilustra os principais componentes de uma extrusora de parafuso único, des-

tacando as suas três secções funcionais, bem como o percurso do material desde a alimentação até à saída pelo molde.

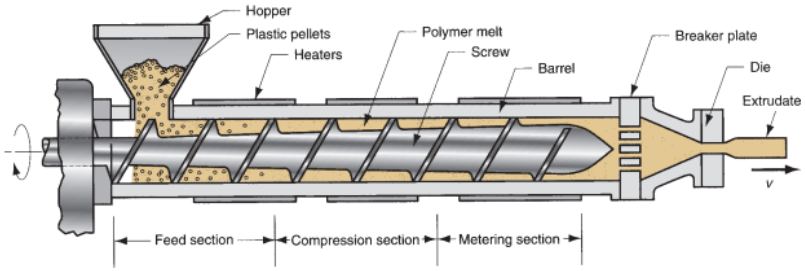


Figura 3.1: Principais componentes de uma extrusora de parafuso único. Fonte: (2)

Para garantir o bom funcionamento deste processo, a máquina está equipada com diversos sensores que monitorizam parâmetros-chave como temperatura, pressão e velocidade. Conforme descrito por Groover (2), a extrusora desempenha várias funções ao longo de três secções principais, cada uma responsável por uma etapa específica do processo:

1. **Secção de alimentação** – Responsável por introduzir o material através do funil de alimentação e iniciar o seu pré-aquecimento.
2. **Secção de compressão** – Transforma o material numa massa fundida, extrai o ar retido entre os grânulos e comprime o material por ação da rotação do parafuso aquecido.
3. **Secção de medição** – Homogeneiza a massa fundida e desenvolve a pressão necessária para forçar o material através do molde.

Vários sensores capturam dados ao longo do processo, permitindo analisar e otimizar o funcionamento da extrusora, reduzir o desperdício e minimizar defeitos (3). As máquinas estão equipadas com quatro sensores de temperatura, um para cada secção do parafuso e um para a temperatura ambiente, dois sensores de pressão, no interior do cilindro e para a pressão ambiente, e um sensor de velocidade para monitorização da rotação do parafuso (3). Cada secção do parafuso é monitorizada individualmente, possibilitando uma análise detalhada do comportamento térmico e da estabilidade operacional do sistema. No total, o sistema integra sete sensores instalados estrategicamente, conforme descrito na Tabela 3.1.

Tabela 3.1: Descrição das variáveis recolhidas no sistema de extrusão.

Campo	**Tipo**	**Descrição**
timestamp	*timestamp*	Instante da medição
ambient_temp	*float*	Temperatura ambiente (°C), entre 18 e 30
*temp*1	*float*	Temperatura na secção de alimentação (°C), entre 130 e 150
*temp*2	*float*	Temperatura na secção de compressão (°C), entre 150 e 180

(continua na próxima página)

Campo	Tipo	Descrição
*temp*3	*float*	Temperatura na secção de medição (°C), entre 180 e 220
ambient_pressure	*float*	Pressão ambiente (hPa), entre 1005 e 1025
pressure	*float*	Pressão no interior do cilindro (bar), entre 70 e 350
rotation	*float*	Velocidade de rotação do parafuso (rpm), entre 20 e 60

Os sensores de temperatura desempenham funções distintas e essenciais em cada secção da extrusora. O sensor instalado na secção de alimentação (*temp*1) monitoriza a temperatura inicial dos grânulos de plástico à medida que estes são transportados do funil para o cilindro (2). Este controlo assegura um pré-aquecimento adequado do material, facilitando a fusão subsequente e prevenindo choques térmicos que poderiam comprometer a qualidade do produto final. Na secção de compressão, o sensor de temperatura (*temp*2) mede a temperatura do material durante o processo de fusão e compressão (2). Esta monitorização permite garantir que o plástico atinge a consistência líquida desejada, ao mesmo tempo que possibilita a extração eficiente do ar retido entre os grânulos, evitando defeitos provocados por bolhas de ar. Já na secção de medição, o sensor (*temp*3) monitoriza a temperatura final do material fundido antes da sua extrusão pela matriz (2). Este controlo é fundamental para assegurar que o material se encontra na temperatura ideal para a moldagem, evitando oscilações na qualidade do produto acabado. Adicionalmente, o sensor de temperatura ambiente (*ambient_temp*) monitoriza as condições térmicas do ambiente de trabalho, permitindo ajustar automaticamente os parâmetros operacionais da extrusora. Este ajuste é crucial, uma vez que a temperatura ambiente pode afetar a eficiência da transferência de calor e, consequentemente, o comportamento do material plástico (3).

Os sensores de pressão desempenham um papel determinante na estabilidade do processo. O sensor instalado no interior do cilindro (*pressure*) monitoriza a pressão exercida durante a extrusão, assegurando que esta se mantém dentro de limites seguros e eficientes para a fusão do material (3). Esta monitorização é essencial para prevenir situações de sobrepressão, que podem comprometer a integridade do equipamento ou afetar a qualidade do produto final. Em contrapartida, o sensor de pressão ambiente (*ambient_pressure*) recolhe dados sobre a pressão atmosférica no ambiente envolvente, permitindo compensar variações externas que possam influenciar o desempenho do processo. A estabilidade da pressão ambiente é particularmente importante, dado que pequenas flutuações podem alterar o comportamento do material plástico e o funcionamento dos sistemas térmicos. É importante notar que estes dois sensores operam com unidades distintas: a pressão no cilindro é medida em bar, enquanto a pressão ambiente é registada em hectopascais (hPa). Esta diferenciação deve-se à escala dos valores envolvidos, a pressão interna pode atingir valores bastante elevados, tornando inadequado o uso de unidades como o hPa, que são mais apropriadas para variações subtis típicas do ambiente atmosférico.

Por outro lado, o sensor de velocidade de rotação (*rotation*) controla a velocidade a que o parafuso da extrusora gira, regulando a taxa de alimentação do material ao longo do cilindro. Esta variável tem um impacto direto na uniformidade da massa fundida e, por consequência, na consistência e qualidade do produto final. Manter a rotação dentro dos parâmetros definidos é, por isso, fundamental para assegurar a eficiência global do sistema

e evitar falhas na produção.

Tal como já foi mencionado, o conjunto de dados considerado neste caso de estudo foi gerado com o objetivo de simular um ambiente industrial realista, incorporando falhas e comportamentos anómalos que ocorrem com frequência em contextos de produção contínua. No total, foram recolhidos 203,273 registos ao longo de três dias consecutivos de operação, com uma frequência de aquisição de um registo por segundo. Os dados simulam medições provenientes de sensores físicos instalados na extrusora, permitindo a monitorização contínua e em tempo real do funcionamento do sistema. Os registos correspondentes aos dois primeiros dias (139,209 registos) foram utilizados como base histórica, de forma a estabelecer referências estatísticas, perfis de comportamento e regras de consistência entre variáveis. O terceiro dia (64,065 registos) serviu como base para a análise em tempo real, permitindo validar as métricas de qualidade definidas e monitorizar desvios operacionais em cenários com maior incerteza. A Figura 3.2, retirada de (3), representa esta amostra, apresentando os valores registados por cada sensor ao longo dos três dias. Esta visão geral do conjunto de dados permite identificar visualmente a ocorrência de valores nulos e anómalos. A amostra em questão foi selecionada precisamente por conter alguns dos problemas de qualidade dos dados mais comuns em ambientes industriais (3).

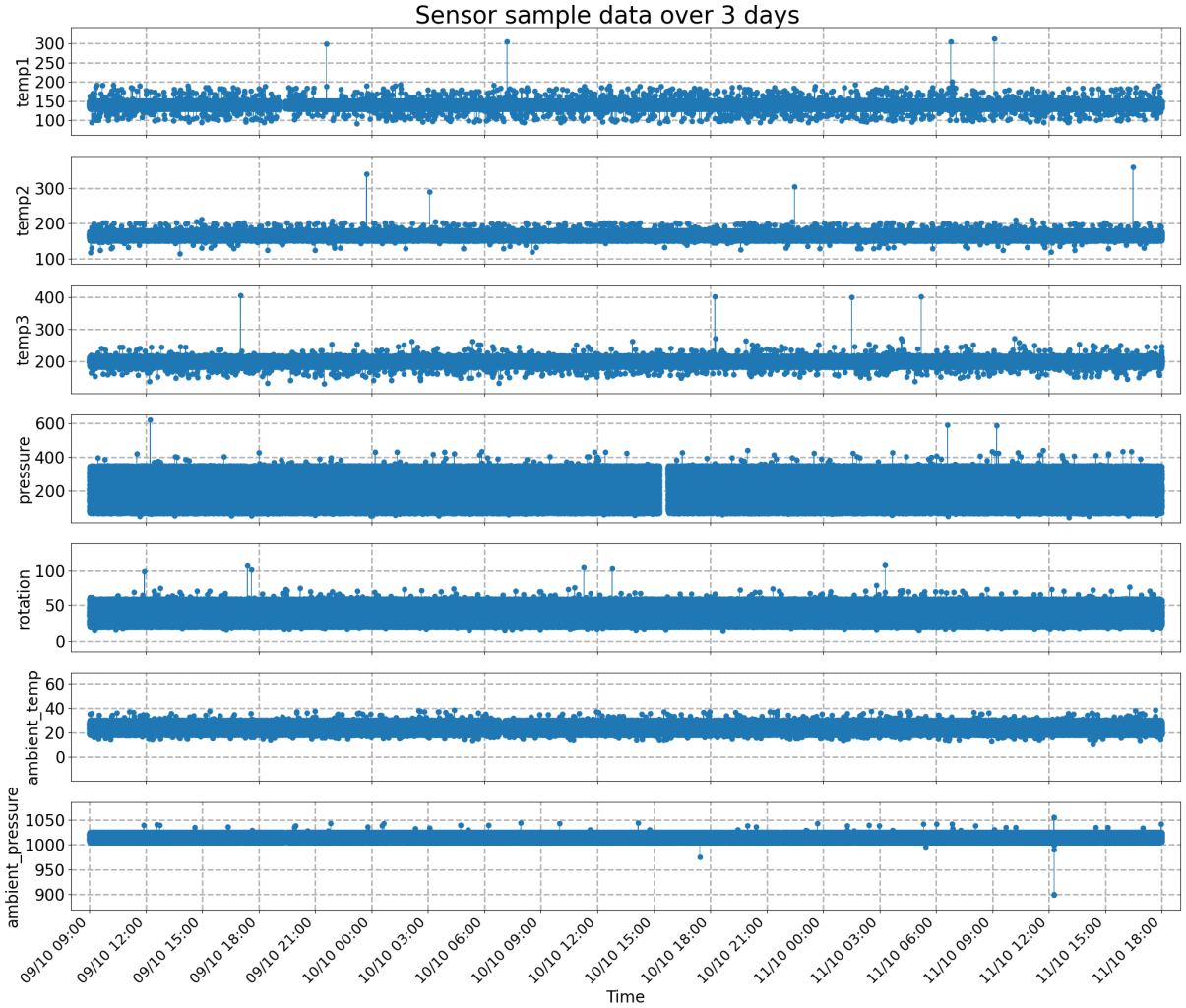


Figura 3.2: Variação de cada sensor ao longo do tempo. Fonte: (3)

Com o objetivo de permitir uma análise contínua e evolutiva, os dados foram organizados

em blocos de cinco minutos. Cada bloco é sujeito a operações de pré-processamento baseadas em *data profiling*, como a identificação de valores nulos, o cálculo de percentis e a verificação de regras de correlação entre variáveis. Estes passos asseguram que as métricas aplicadas são ajustadas ao contexto temporal e operacional do processo, possibilitando uma avaliação mais robusta da qualidade dos dados. Esta abordagem permitiu não só monitorizar a qualidade dos dados ao longo do tempo, como também identificar relações entre deteriorações na qualidade e comportamentos atípicos dos sensores. O caso de estudo foi, assim, concebido para avaliar de forma sistemática quatro dimensões fundamentais da qualidade dos dados, acurácia, completude, consistência e atualidade, e explorar o seu contributo para a deteção de eventos críticos em ambientes de I4.0.

Com base na estrutura descrita e nas características do conjunto de dados, foram aplicadas métricas específicas para cada uma das quatro dimensões selecionadas da qualidade dos dados. Estas métricas foram escolhidas com base na literatura e na avaliação feita por Peixoto et al. em (1; 3). A análise foi conduzida de forma incremental, permitindo acompanhar a evolução da qualidade ao longo do tempo e detetar alterações relevantes no comportamento do sistema.

Nas secções seguintes, descreve-se em detalhe a metodologia utilizada para o cálculo de cada métrica, bem como os resultados obtidos, destacando os contributos de cada dimensão para a compreensão da fiabilidade dos dados e para o suporte à deteção precoce de falhas em ambientes industriais.

3.1.1 Acurácia

Para avaliar a dimensão da acurácia, foi aplicada a métrica descrita na equação 2.2. Esta abordagem, inspirada nos trabalhos de Goknil et al. (9) e Sicari et al. (52), permite calcular a acurácia média para cada variável numérica em cada bloco j de cinco minutos. A métrica normaliza os valores dentro de um intervalo de $[0, 1]$, onde resultados próximos de 0 indicam observações junto do valor mínimo aceitável, enquanto valores próximos de 1 indicam proximidade ao valor máximo.

Na métrica 2.2, X representa o conjunto de todas as observações disponíveis até ao bloco j, funcionando como histórico de referência. Para mitigar o impacto de outliers, os limites inferior e superior do intervalo ($\max(X)$ e $\min(X)$) são definidos, não de forma absoluta, mas com base nos percentis 10 e 90 do histórico. Esta adaptação, proposta em (3), constitui uma extensão à abordagem de Goknil, conferindo maior robustez e sensibilidade ao comportamento real do processo industrial, frequentemente sujeito a variações operacionais legítimas que não devem ser tratadas como anomalias. Com base na análise e em iterações experimentais, os percentis 10 e 90 foram escolhidos como limites de referência por se revelarem mais eficazes na representação dos valores mínimos e máximos aceitáveis para cada variável. A Figura 3.3, retirada de (3), ilustra essa diferença: demonstra a variação do valor mínimo e máximo real registado em cada bloco de cinco minutos (linhas tracejadas verde e vermelha), juntamente com os limites dinâmicos definidos pelos percentis 10 e 90 (linhas contínuas azul e laranja). Observa-se que os valores extremos reais flutuam consideravelmente ao longo do tempo, o que comprometeria a consistência da métrica de acurácia caso fossem utilizados diretamente. Os percentis, por outro lado, oferecem uma alternativa mais estável e resiliente a desvios pontuais, garantindo limites de normalização mais fiáveis e adequados à evolução do processo.

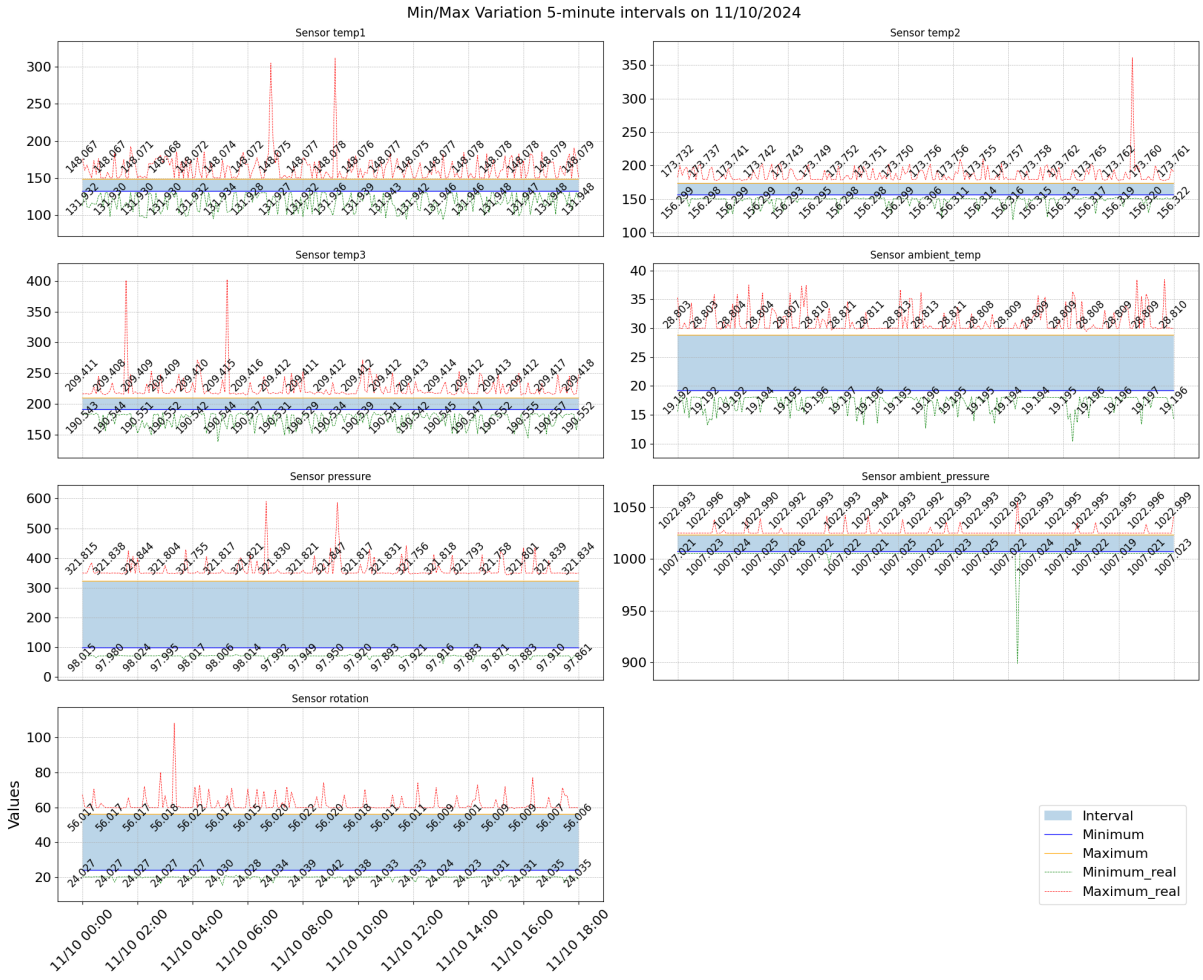


Figura 3.3: Variação dos valores mínimo e máximo por bloco, comparando com os percentis 10 e 90. Fonte: (3)

Quando o resultado da acurácia de um bloco j se mantém dentro do intervalo $[0, 1]$, infere-se que os dados presentes nesse bloco seguem um padrão aceitável. Por outro lado, valores fora desse intervalo indicam desvios significativos face ao comportamento histórico, podendo refletir imprecisões relevantes ou falhas de medição. Neste sentido, a métrica de acurácia assume um papel importante como indicador para a deteção de anomalias nos dados.

Os resultados obtidos para esta métrica, aplicados ao conjunto de dados ilustrado na Figura 3.2, são apresentados na Figura 3.4. Esta figura mostra a evolução da acurácia ao longo do tempo para todos os sensores, permitindo identificar entre quatro a cinco situações no dia 11 de outubro em que os valores de acurácia se aproximam dos limites do intervalo $[0, 1]$.

Como os dados são analisados em blocos de cinco minutos (cerca de 300 registos por bloco), a acurácia média de cada bloco é calculada dinamicamente com base na normalização entre os percentis 10 e 90, o que permite acompanhar flutuações em tempo real. Espera-se, assim, que a maioria das leituras normalizadas se concentre numa faixa central, geralmente entre [0,4, 0,6]. Desvios em relação a este intervalo indicam a presença de valores atípicos

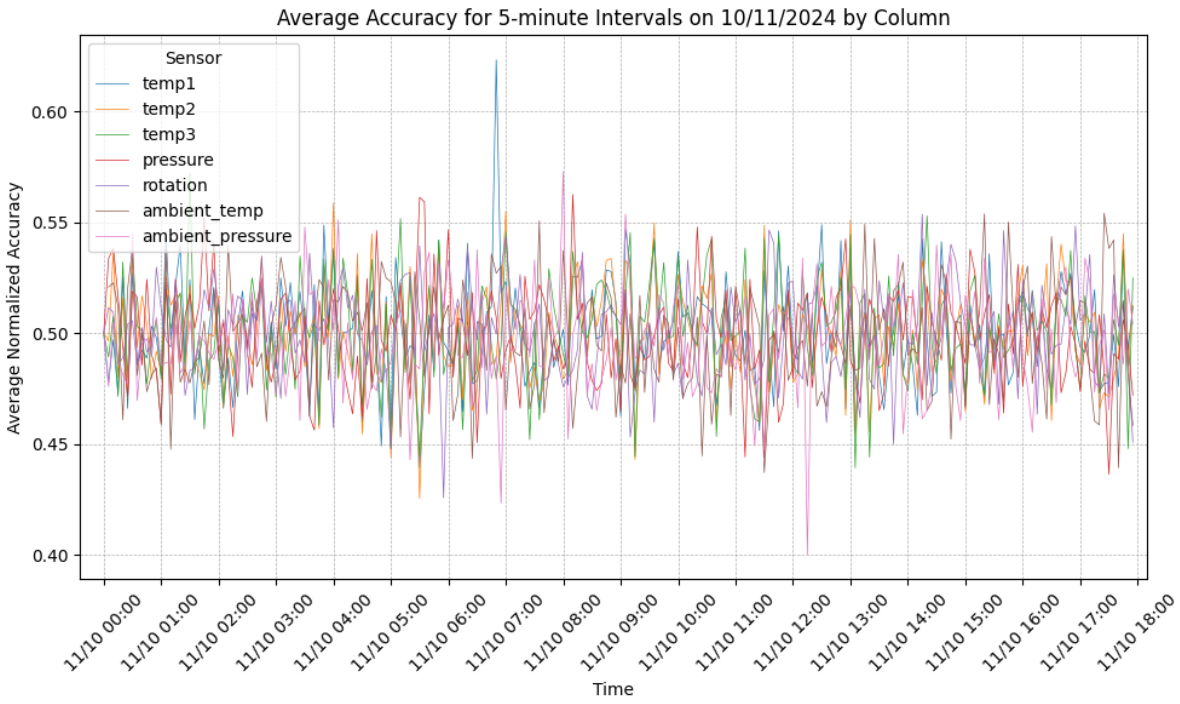


Figura 3.4: Resultados da métrica da acurácia. Fonte: (3)

ou inconsistentes no bloco com o comportamento habitual do sistema, podendo sinalizar anomalias nos sensores ou alterações inesperadas no processo industrial. Adicionalmente, os autores em (3) apresentam uma visualização complementar, onde os mesmos dados são apresentados por sensor, em gráficos individuais. Esta representação facilita a análise detalhada de cada variável, permitindo isolar com maior precisão os momentos em que ocorreram quebras na acurácia e identificar possíveis causas associadas.

Para o cáculo da métrica 2.2, foram utilizadas tarefas de *data profiling*, como a identificação de valores nulos, extremos e tipos de dados em cada coluna. Estas tarefas correspondem às categorias de *data profiling* identificadas por: cardinalidade, distribuições de valores e tipos de dados, padrões e domínios.

3.1.2 Completude

Em ambientes de I4.0, lacunas nos dados podem comprometer a monitorização contínua do sistema e limitar a fiabilidade de análises posteriores, pelo que é essencial identificar tanto a ausência de dados a nível do elemento e do registo, como a verificação da recolha efetiva dos valores esperados. Para esta dimensão foram consideradas duas métricas distintas. A primeira, expressa na equação 2.3, calcula a razão entre os valores efetivamente disponíveis e os que eram esperados. Embora esta métrica possa ser aplicada tanto ao nível das colunas como das linhas, neste caso optou-se por analisar os resultados ao nível da linha, de forma a avaliar a completude de cada registo individual. As tarefas de *data profiling* utilizadas incluíram a contagem de valores nulos e válidos por atributo e por registo, bem como a identificação do tipo de dados de cada campo, enquadrando-se nas categorias de cardinalidade e de tipos de dados, padrões e domínios.

A Figura 3.5 de (3) apresenta os resultados da métrica ao nível da linha, onde se observa

que a completude se mantém, na maioria dos blocos, no valor ideal de 1. No entanto, entre as 5:55 e as 6:05, há uma quebra notável, com a completude a descer para 0,79 e 0,66 em dois blocos consecutivos. Estes valores sugerem ausência de dados em múltiplos campos e levantam suspeitas de falha momentânea de comunicação com os sensores ou problemas de transmissão de dados.

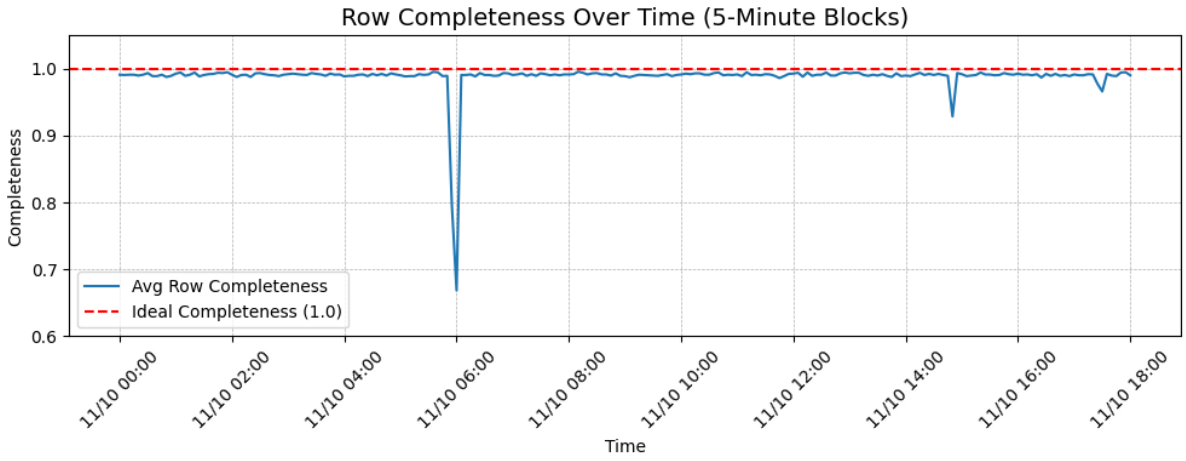


Figura 3.5: Completude por linha em blocos de 5 minutos. Fonte: (3)

Adicionalmente, foi também utilizada a métrica 2.4 que avalia a completude com base na regularidade temporal da chegada dos dados. Esta métrica compara o número real de ocorrências num intervalo com o número de ocorrências esperadas, permitindo detetar lacunas de registos a partir dos *timestamps* e utiliza as tarefas contagem do número de linhas na categoria cardinalidade e a tarefa de identificação de tipos de dados na categoria tipos de dados, padrões e domínio de *data profiling*. A Figura 3.6, também retirada de (3), apresenta os resultados desta métrica ao longo do dia 11 de outubro, evidenciando que, embora a maioria dos blocos mantenha um valor próximo de 1, existem períodos onde a completude diminui, refletindo possíveis falhas na recolha dos dados. Nenhum dos blocos analisados registou um valor de completude superior ao esperado, o que, caso ocorresse, poderia indicar duplicação de registos.

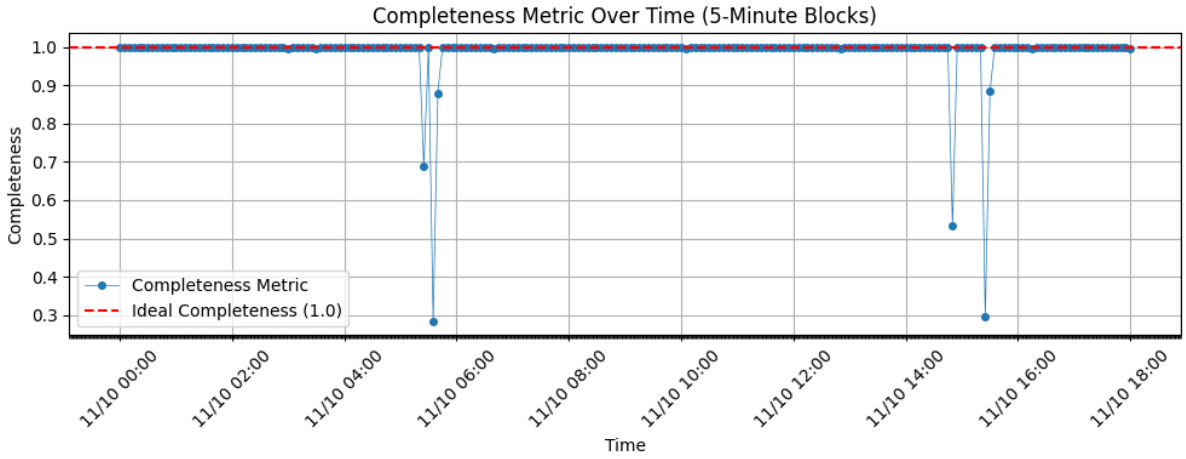


Figura 3.6: Completude global por bloco (linha azul) e valor ideal (linha a 1). Fonte: (3)

Deste modo, estas duas métricas complementares permitem avaliar a completude de forma

multidimensional: por um lado, identificam falhas associadas à ausência de atributos específicos, por outro, revelam perdas globais de registos ao longo do tempo. Juntas, constituem uma ferramenta valiosa para a deteção de falhas operacionais em tempo real e para a melhoria da qualidade da monitorização em sistemas industriais.

3.1.3 Consistência

A dimensão da consistência avalia a conformidade dos dados com regras lógicas, dependências funcionais ou relações esperadas entre variáveis. Em ambientes industriais, esta dimensão é particularmente relevante para detetar incoerências que possam resultar de falhas nos sensores, erros de medição ou problemas de integração entre sistemas. Para o cálculo da consistência utilizou-se apenas a métrica 2.5 (3), sendo necessário recorrer ao conjunto de dados históricos para estabelecer regras de referência com base na identificação de padrões de correlação entre variáveis.

A identificação das regras entre colunas foi realizada com base na análise de correlações, recorrendo a tarefas de *data profiling*. Foi definido um limiar de 0,7 como critério mínimo para considerar a existência de uma regra com forte probabilidade de associação. Embora a correlação, por si só, não comprove uma dependência funcional, permite sinalizar relações relevantes entre sensores (3). Importa referir que as regras devem ser validadas ou definidas por especialistas com conhecimento do domínio.

Com base nesta análise, Peixoto et al. (3), identificaram três correlações fortes entre sensores de temperatura: entre *temp*1 e *temp*2, entre *temp*2 e *temp*3, e entre *temp*1 e *temp*3. A Figura 3.7 apresenta os resultados da métrica de consistência ao longo do dia 11 de outubro. Observa-se que, na maioria dos blocos de cinco minutos, o valor da consistência é 1, indicando que todas as três regras foram verificadas. Em alguns blocos, a métrica assume o valor de 0,66, indicando que apenas duas das três relações esperadas foram observadas. Num bloco isolado, o valor da métrica desce para 0,33, sinalizando a verificação de apenas uma das regras, o que pode refletir um comportamento atípico ou uma falha pontual em um dos sensores envolvidos.

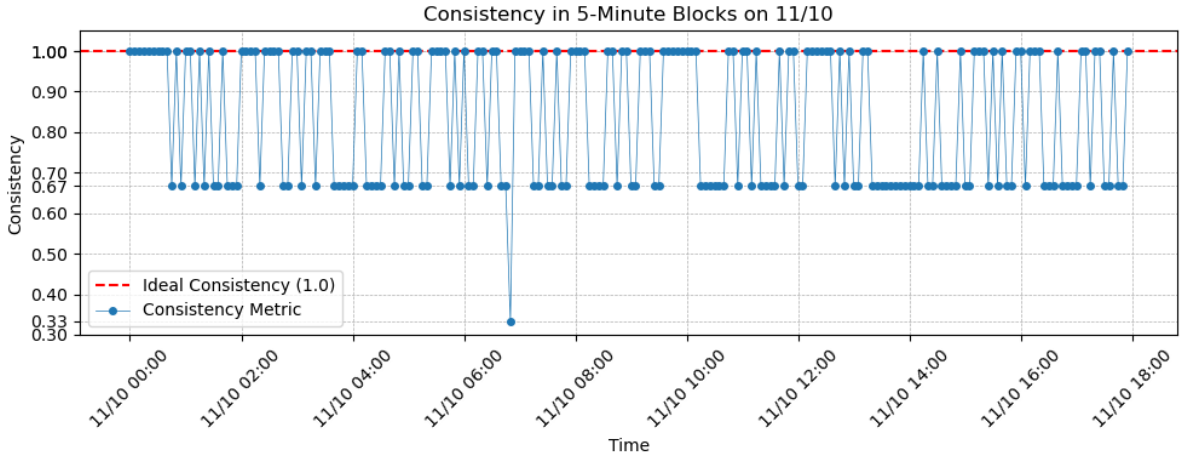


Figura 3.7: Resultados da métrica de consistência ao longo do tempo. Fonte: (3)

A métrica de consistência constitui, assim, um mecanismo complementar às restantes dimensões, contribuindo para a deteção de anomalias resultantes da violação de padrões esperados entre variáveis interdependentes.

3.1.4 Atualidade

A dimensão da atualidade refere-se à relevância temporal dos dados, isto é, ao grau em que estes permanecem válidos para o momento em que são analisados. Em ambientes industriais com monitorização contínua, é fundamental garantir que os dados utilizados na tomada de decisão reflitam com precisão o estado atual do sistema.

A métrica utilizada por Peixoto et al. (3) para avaliar esta dimensão é a métrica 2.9, esta baseia-se na idade dos dados (*currency*), definida como o intervalo de tempo entre o momento da recolha e o momento atual da análise. Esta idade é obtida a partir do campo *timestamp*, identificado por técnicas de *data profiling* na categoria de tipos de dados, padrões e domínios. Um parâmetro essencial nesta métrica é a volatilidade, que define a janela temporal durante a qual os dados são considerados relevantes. Este intervalo pode ser definido com base no conhecimento do domínio, normas operacionais ou experiência prática (3). Neste caso de estudo, foi assumido um valor de volatilidade de 10 minutos, significando que os dados são considerados válidos durante esse período a partir do instante da sua recolha. Importa referir que este valor pode variar conforme as características do processo monitorizado ou o tipo de sensor utilizado.

A métrica de atualidade resulta, assim, num valor normalizado entre 0 e 1, em que 1 representa dados perfeitamente atuais e 0 indica que os dados ultrapassaram o limite de validade temporal definido. Quanto mais próximo de 1, maior é a relevância do dado para a análise em tempo real, reforçando a sua utilidade para a tomada de decisão. Nos resultados obtidos verifica-se que, durante grande parte do dia, a atualidade permaneceu nula (valor 0), refletindo o facto de os dados analisados já se encontrarem fora da janela de volatilidade estabelecida. No entanto, nos blocos finais (entre as 17:50 e as 18:00), observou-se um aumento progressivo dos valores da métrica, atingindo 0,25 e 0,7499, resultado da análise coincidir com os registos mais recentes do conjunto de dados.

Esta abordagem à atualidade permite incorporar diretamente a dimensão temporal na avaliação da qualidade dos dados, sendo particularmente útil em sistemas com elevada frequência de atualização ou requisitos de resposta em tempo real.

3.1.5 Resultados

Os resultados obtidos demonstram a aplicabilidade das métricas propostas em cenários industriais e evidenciam a sua eficácia na identificação de diferentes tipos de problemas associados à qualidade dos dados. A análise efetuada ao longo de blocos temporais de cinco minutos permitiu uma deteção localizada de desvios e anomalias, possibilitando uma leitura mais granular do comportamento dos sensores ao longo do tempo. Esta abordagem revelou-se especialmente útil para identificar variações súbitas ou padrões de falha com impacto limitado no tempo, que poderiam passar despercebidos.

A integração das métricas com tarefas de *data profiling* mostrou-se essencial para garantir o seu cálculo eficiente e adaptado ao contexto dinâmico dos dados industriais. As tarefas de *data profiling*, como a contagem de valores nulos, a análise de tipos de dados ou até a identificação de dependências entre atributos, contribuíram diretamente para o cálculo das métricas. Adicionalmente, os autores em (3) apresentam uma figura de síntese que evidencia de forma estruturada a correspondência entre os diferentes elementos do processo de *data profiling* e as DQD avaliadas neste estudo. Esta figura apresenta uma cadeia lógica de mapeamento, que se inicia nos tipos de análise, passa pelas categorias de

tarefas de *data profiling*, associa cada tarefa a uma métrica concreta, e finaliza nas DQD que essas métricas avaliam. Esta representação reforça o papel do *data profiling* como componente central e estruturante na avaliação sistemática da qualidade dos dados.

A variabilidade observada entre sensores reforça a importância de realizar avaliações diferenciadas. Fatores como a sensibilidade do sensor, a estabilidade do processo ou o posicionamento físico no sistema influenciam diretamente os indicadores de qualidade, tornando necessária uma análise contextualizada e orientada ao domínio.

Este primeiro caso de estudo valida a fase inicial do *pipeline* proposto, demonstrando a viabilidade da aplicação de métricas baseadas em *data profiling* para ambientes de produção contínua. Com base nos resultados obtidos, foi possível avançar para uma extensão desta análise, centrada na deteção de anomalias na mesma fonte de dados, recorrendo a técnicas estatísticas e de aprendizagem automática. Esta extensão é apresentada na subsecção seguinte.

3.1.6 Análise Comparativa de Técnicas de Deteção de Anomalias

Dando continuidade ao primeiro caso de estudo, foi realizada uma análise complementar focada na deteção de anomalias em séries temporais, utilizando os dados recolhidos entre as 9h e as 12h do sensor *temp*1. Esta extensão foi desenvolvida no âmbito de um artigo apresentado na conferência *Business and Technology 2025*[2]. O principal objetivo foi comparar a eficácia de diferentes técnicas na identificação de valores atípicos, avaliando a sua aplicabilidade em contextos industriais de monitorização em tempo real.

A deteção de anomalias é uma tarefa crítica em vários domínios, nomeadamente na manutenção preventiva, na deteção de falhas, na prevenção de fraudes e na supervisão de sistemas (31; 32). Em ambientes de IdC, a elevada volumetria de dados e a necessidade de processamento em tempo real tornam necessário encontrar um equilíbrio entre desempenho computacional e sensibilidade na deteção. Este desafio é agravado pela necessidade de compreender a distribuição dos dados e de identificar desvios subtis ou padrões irregulares, exigindo abordagens adaptadas à natureza do problema e ao contexto operacional.

Neste estudo, foram analisadas quatro abordagens distintas: (i) a métrica de acurácia (Eq. 2.2), (ii) o algoritmo probabilístico *t-digest*, (iii) a decomposição da série temporal com STL, e (iv) o modelo LSTM, representativo de técnicas de aprendizagem profunda. Cada método foi avaliado com base no número de anomalias detetadas e no tempo de execução.

A primeira abordagem consistiu na aplicação da métrica de acurácia descrita na equação 2.2, com o objetivo de identificar valores extremos através da normalização dos dados. Inicialmente, foram utilizados os percentis 10 e 90 como limites de referência para esta normalização, tal como descrito por Peixoto et al. (3). No entanto, uma análise mais detalhada revelou que esta escolha não era suficientemente eficaz para séries temporais com variação. Esta abordagem revelou-se pouco adaptativa, resultando numa deteção artificialmente simétrica de anomalias, onde foram detetados exatamente 898 valores abaixo do limiar inferior e 898 acima do superior, totalizando 1796 anomalias. Estes resultados

[2]BTC 2025: `https://businesstechnologyconference.com/`

indicam que a métrica estava a classificar como anómalos muitos valores que, na realidade, se enquadravam dentro do padrão esperado, o que compromete a fiabilidade do processo. Para mitigar esta limitação, foi adotada uma estratégia alternativa baseada na regra empírica dos 68–95–99.7%, segundo a qual aproximadamente 99.73% dos dados se encontram dentro de três desvios padrão da média. Esta regra, amplamente utilizada na estatística para identificar valores atípicos, permitiu definir limites mais rigorosos e mais adequados (71). Assim, foram utilizados os quantis 0,00135 e 0,99865 como novos valores mínimo e máximo para a normalização da métrica da acurácia, conforme proposto por Sterjev (72). Esta reformulação aumentou a sensibilidade da métrica a desvios significativos, reduzindo os falsos positivos e proporcionando uma deteção mais realista e eficaz de anomalias.

Em seguida, foi aplicado o algoritmo *t-digest*, uma estrutura probabilística eficiente para estimar estatísticas de ordem, como medianas e percentis, destacando-se pela sua elevada precisão e baixo consumo de memória (37). A estratégia adotada consistiu em utilizar o *t-digest* para estimar os quantis de referência a aplicar na métrica de acurácia, replicando a mesma lógica da normalização com base nos quantis 0,00135 e 0,99865. Todos os valores situados fora deste intervalo foram classificados como *outliers*. Embora partilhe os mesmos limites teóricos da métrica, o *t-digest* apresenta a vantagem de adaptar os cálculos à distribuição real dos dados, oferecendo maior flexibilidade face a variações no comportamento do processo. Esta combinação de adaptabilidade e eficiência torna-o particularmente adequado para contextos industriais em tempo real, onde é fundamental garantir um equilíbrio entre desempenho computacional e sensibilidade na deteção de anomalias.

A terceira técnica explorada baseou-se na decomposição da série temporal com o método STL, que permite separar a tendência, a sazonalidade e os resíduos da série (42). Neste estudo, a deteção de anomalias foi realizada aplicando limiares aos valores residuais obtidos na decomposição. De acordo com a abordagem proposta por Sterjev (72), todos os resíduos que ultrapassavam o intervalo definido por quantis 0,00135 e 0,99865 foram classificados como anómalos. Esta técnica permite isolar desvios não explicáveis pelas variações estruturais do processo, fornecendo uma perspetiva complementar à análise baseada em limites absolutos.

Por fim, foi implementado um modelo LSTM, uma arquitetura de rede neural recorrente concebida para lidar com dependências de longo prazo em séries temporais (44). A deteção de anomalias foi efetuada com base no erro absoluto entre os valores reais e as previsões geradas pelo modelo. Seguindo a abordagem proposta por Ho (73), considerou-se anómala qualquer observação cujo erro ultrapassasse um limiar definido a partir da distribuição dos erros mais frequentes.

As quatro abordagens foram aplicadas de forma sistemática ao mesmo subconjunto de dados, permitindo uma comparação direta dos seus pontos fortes e limitações. A análise centrou-se na relação entre o tempo de execução e a sensibilidade na deteção de anomalias. A Tabela 3.2 apresenta um resumo dos principais resultados obtidos para cada método. Para complementar a análise, foi disponibilizado um repositório *GitHub* com as visualizações correspondentes a cada técnica aplicada, facilitando a comparação visual do desempenho e dos padrões de deteção observados (74).

Tabela 3.2: Resultados da aplicação de técnicas de deteção de anomalias.

Técnica	**Tempo de Execução (s)**	**Nº de Anomalias**
Métrica da Acurácia	0.0258	26
T-Digest	0.1954	24
STL	1.1059	26
LSTM	2998.8978	139

As abordagens mais rápidas, a métrica de acurácia e o algoritmo *t-digest*, demonstraram elevada eficiência na deteção de anomalias mais evidentes, revelando-se particularmente adequadas para uma fase rápida e direta de identificação de anomalias graves em tempo real, onde a velocidade de processamento é um requisito crítico. Por outro lado, o modelo LSTM apresentou uma sensibilidade significativamente superior, detetando um número muito mais elevado de anomalias, embora com um tempo de execução substancialmente superior. Este comportamento torna-o mais apropriado para contextos de análise aprofundada, onde a precisão na deteção de padrões complexos justifica o custo computacional acrescido.

As três primeiras abordagens (acurácia, *t-digest* e STL) detetaram um número semelhante de anomalias, embora com pequenas variações nos registos específicos identificados como fora do padrão. As imagens disponibilizadas no repositório *GitHub* ilustram essas diferenças, evidenciando que, apesar da semelhança quantitativa, os métodos não coincidem completamente na deteção dos mesmos eventos. Em particular, os resultados do *t-digest* e do STL estão bastante alinhados com os da métrica de acurácia, mas diferem quanto à complexidade computacional: o *t-digest* oferece uma execução mais eficiente, enquanto o STL, embora mais exigente em termos de tempo, acrescenta valor pela sua capacidade de separar padrões sazonais de desvios inesperados. O modelo LSTM, por sua vez, destacou-se pela sua elevada sensibilidade, sendo capaz de identificar um número muito maior de anomalias, muitas das quais não captadas pelas restantes técnicas. Esta capacidade de deteção fina pode ser particularmente útil em cenários que exigem uma análise preditiva detalhada, nomeadamente na antecipação de falhas subtis ou alterações progressivas no comportamento do sistema.

Em síntese, os resultados obtidos validam a relevância da métrica de acurácia como uma abordagem eficaz para a deteção de anomalias em séries temporais industriais. Embora inicialmente baseada em percentis dinâmicos, a métrica revelou ainda melhor desempenho quando ajustada com base na regra estatística dos 99.73%, aumentando significativamente a sua sensibilidade e capacidade de discriminação de valores atípicos. Este estudo complementar reforça, assim, o valor das métricas baseadas em *data profiling* na monitorização da qualidade dos dados e abre caminho à sua integração com técnicas mais avançadas nos próximos casos de estudo.

3.2 Caso de Estudo 2 - Índice da Qualidade

Este estudo incide sobre um sistema de bombagem de água pertencente a uma pequena comunidade rural, caracterizada por falhas técnicas recorrentes que originam interrupções

no abastecimento de água. Estas falhas têm repercussões diretas na vida quotidiana da população local e na produtividade agrícola e industrial da região, tornando este tipo de infraestrutura crítica para o funcionamento regular da comunidade (75). Em contextos mais isolados, a obsolescência tecnológica e a manutenção deficiente acentuam a vulnerabilidade do sistema, afetando a fiabilidade dos serviços prestados e, por conseguinte, a qualidade de vida das populações (4).

Este caso de estudo baseia-se no artigo ***Data Quality Assessment: A Practical Application*** de Costa e Silva et al. (4) e foi desenvolvido numa fase intermédia do projeto. Corresponde à transição entre uma aplicação inicial das métricas de qualidade dos dados e a análise comparativa entre diferentes técnicas de deteção de anomalias que levou a uma reformulação da métrica da acurácia (ver Caso de Estudo 1 e Eq. 2.2).

Os dados utilizados provêm de um repositório público[3] e dizem respeito ao funcionamento de um sistema de bombagem entre abril e agosto de 2018. A rede de monitorização inclui 52 sensores que recolhem variáveis como pressão, velocidade, temperatura e humidade (75) com uma frequência de um minuto, resultando num elevado volume de dados contínuos adequados à análise de padrões e identificação de falhas (4).

Para esta análise, foi selecionado o mês de abril, totalizando 43200 registos. A Figura 3.8 ilustra os dados de quatro sensores nesse período[4]. Dois eventos de falha estão documentados: o primeiro entre os dias 12 e 13 e o segundo entre os dias 18 e 20 de abril, ambos marcados por interrupções prolongadas no serviço.

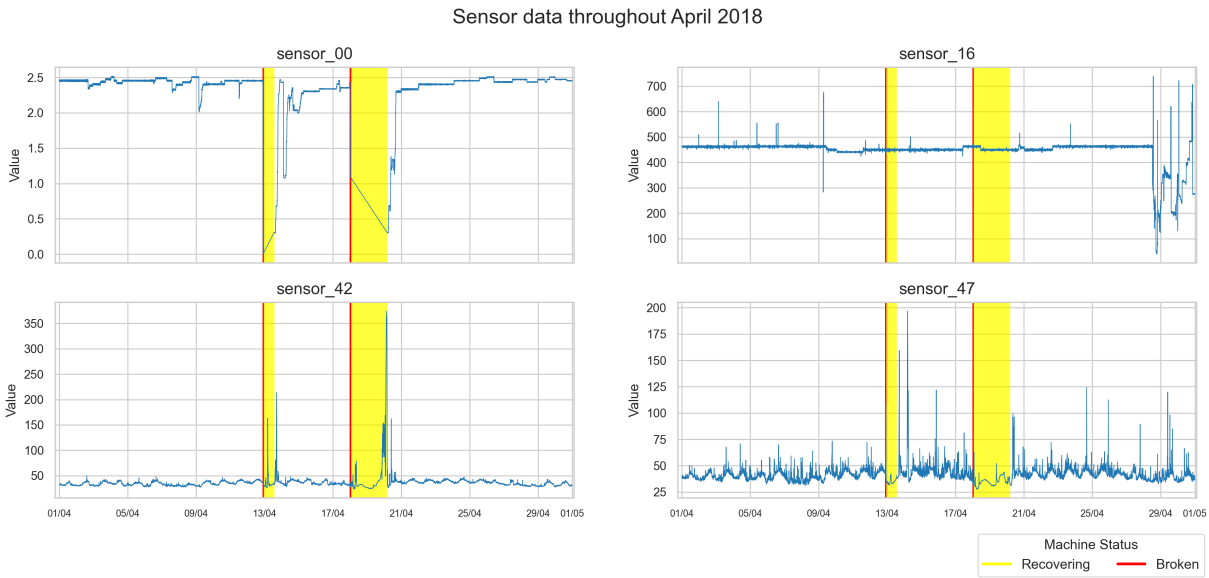


Figura 3.8: Dados de 4 dos 52 sensores, ao longo de abril de 2018. Fonte: (4)

A seguir, apresentam-se as fórmulas correspondentes ao cálculo dos índices *Quality Score Delta* (QSD), *Weighted Quality Score* (WQS) e *Longitudinal Weighted Quality Score* (LWQS) bem como a forma como foram aplicados ao conjunto de dados selecionado.

[3] `https://www.kaggle.com/datasets/nphantawee/pump-sensor-data`

[4] Todos os gráficos dos sensores em `https://github.com/TeresaPeixoto/Data-Quality-Assessment`https://github.com/TeresaPeixoto/Data-Quality-Assessment (74)

3.2.1 *Quality Score Delta*

A necessidade de uma avaliação abrangente e contínua da qualidade dos dados em ambientes industriais motivou o desenvolvimento de um índice composto, designado por QSD. A abordagem proposta visa apoiar a tomada de decisão com base em métricas objetivas e consistentes, desenhadas do ponto de vista do utilizador final (4). Ao consolidar avaliações dispersas numa métrica composta e personalizável, facilita-se a deteção de degradações, a comparação entre blocos temporais e a definição de estratégias de melhoria contínua na gestão da qualidade dos dados. Este indicador permite capturar, de forma adaptativa e orientada à decisão, variações na qualidade dos dados ao longo do tempo, potenciando a antecipação de anomalias e a deteção precoce de falhas (4).

Diferenciando-se das abordagens convencionais que avaliam cada métrica de forma isolada, o QSD integra dinamicamente os contributos considerados mais relevantes, conciliando uma análise instantânea com uma perspetiva histórica da evolução da qualidade. Esta proposta, inspirada em trabalhos prévios aplicados ao domínio da extrusão (76), foi generalizada para múltiplos contextos industriais, eliminando a dependência de cenários específicos e reduzindo a subjetividade na interpretação dos resultados (4).

A metodologia baseia-se na segmentação temporal dos dados em blocos de 5 minutos, permitindo uma avaliação granular e atualizada. Algumas métricas utilizadas consideram apenas os dados do bloco atual, enquanto outras incorporam também a informação histórica recente, reforçando a robustez e a estabilidade da análise (3). O cálculo do QSD assenta em dois componentes principais: o WQS, que fornece uma avaliação da qualidade do bloco atual e o LWQS, que incorpora dados dos blocos anteriores, atribuindo maior peso aos mais recentes, através de uma função de decaimento exponencial. Esta abordagem dual permite capturar não apenas variações abruptas, mas também tendências acumuladas ao longo do tempo, sendo especialmente relevante em sistemas de monitorização contínua (4).

Um dos principais diferenciais do índice proposto reside na sua capacidade de personalização, permitindo ajustar os pesos atribuídos a cada dimensão e a cada sensor, de acordo com o conhecimento do domínio ou os objetivos operacionais do sistema. No presente estudo, foram consideradas duas dimensões: acurácia (métrica 2.2) e completude (métrica 2.3 ao nível do registo). As dimensões de consistência e atualidade foram propositadamente excluídas, uma vez que os dados provêm de uma rede sensorial homogénea, minimizando o risco de inconsistências estruturais ou semânticas e a ingestão de dados ocorre com baixa latência e em janelas reduzidas, não representando a atualidade um fator crítico neste contexto (4). Assim, optou-se por centrar a análise naquilo que mais afeta a fiabilidade dos diagnósticos: a precisão dos valores registados e a sua presença contínua. A avaliação detalhada dos resultados obtidos para cada uma das duas dimensões está disponível em (4), onde são apresentados os gráficos por sensor e por registo ao longo do mês de abril.

A métrica QSD é particularmente sensível a variações abruptas, funcionando como um mecanismo de alerta precoce. Para cada bloco de dados j, o índice é calculado de acordo com a equação 3.1:

$$QSD_j = \begin{cases} WQS_j - LWQS_j & \text{if } j > 1 \\ 1 & \text{if } j = 1 \end{cases} \tag{3.1}$$

Valores positivos indicam uma melhoria da qualidade relativamente à média recente, enquanto valores negativos sinalizam uma possível degradação súbita que deve ser monitorizada com atenção. Tanto o WQS como o LWQS integram as dimensões de acurácia e completude, ponderadas através dos pesos w_a e w_c, respetivamente, sendo que $w_a + w_c = 1$ e $w_a, w_c \geq 0$. Neste caso, os autores atribuíram maior peso à completude ($w_c = 0.6$) face à acurácia ($w_a = 0.4$), refletindo a prioridade em garantir a presença contínua de dados num cenário de monitorização em tempo real (4). Ambos os componentes permitem ainda a ponderação individual dos sensores, ajustando a sensibilidade do índice às variáveis mais críticas. No presente estudo, todos os sensores foram ponderados de forma equitativa ($w_s = 0.019$), exceto o último, cuja ponderação foi ajustada para garantir que a soma total dos pesos seja unitária (4). Esta estratégia pode, e deve, ser adaptada para refletir a importância relativa de determinados sensores em cenários reais.

Antes de apresentar o WQS e o LWQS, clarifica-se a nomenclatura utilizada nas respetivas fórmulas:

- $s \in S = \{1, ..., n_s\}$ representa o sensor,
- n_j é o número de linhas do bloco j,
- a_j^s é o número de linhas com valores considerados precisos para o sensor s no bloco j,
- c_j^s é o número de linhas sem valores em falta para o sensor s,
- w_s são os pesos de cada sensor s, tal que:
  a) $w_1 + ... + w_{n_s} = 1$ (soma total dos pesos dos sensores),
  b) $w_1, ..., w_{n_s} \geq 0$, para todo $s \in S$.
- $K = \{j - m, ..., j - 1\}$ é o conjunto de números inteiros que contém os índices dos últimos m blocos de dados antes do bloco j
- f_k é uma função de ponderação definida como $f(k) = \exp\left(-\frac{j-k-1}{\beta}\right)$ onde $\beta > 0$ controla a taxa de decaimento e $m > 0$ define o intervalo temporal considerado. Esta função atribui maior peso aos blocos mais recentes.

3.2.2 *Weighted Quality Score*

WQS fornece uma avaliação instantânea da qualidade dos dados recolhidos num determinado bloco temporal j, permitindo identificar, de forma agregada, a fiabilidade dos registos obtidos nesse intervalo. A sua formulação incorpora duas DQD, acurácia e completude, ponderadas de acordo com a sua importância relativa (w_a e w_c), bem como a contribuição específica de cada sensor, através de um conjunto de pesos (w_s). Esta abordagem assegura uma avaliação flexível e adaptável, sensível tanto às características dos dados como às prioridades do sistema monitorizado.

A equação 3.2 define o cálculo do WQS para o bloco j:

$$WQS_j = w_a \left(\sum_{s \in S} w_s \frac{a_j^s}{n_j} \right) + w_c \left(\sum_{s \in S} w_s \frac{c_j^s}{n_j} \right) \tag{3.2}$$

A estrutura da fórmula garante que cada dimensão contribui proporcionalmente à sua importância relativa e ao comportamento dos sensores associados. Esta métrica é especialmente útil em sistemas com fluxos contínuos de dados, pois oferece uma leitura rápida do estado atual da qualidade, podendo ser utilizada para alertas imediatos ou análises operacionais em tempo real.

3.2.3 *Longitudinal Weighted Quality Score*

Complementando a análise pontual da qualidade dos dados providenciada pelo WQS, o LWQS introduz uma componente histórica que permite avaliar a evolução temporal da qualidade dos dados. Esta métrica baseia-se numa janela deslizante composta por múltiplos blocos anteriores ao instante atual, atribuindo maior peso aos blocos mais recentes através de uma função de decaimento exponencial.

Esta abordagem permite capturar padrões de degradação gradual, suavizar flutuações esporádicas e identificar tendências consistentes ao longo do tempo. Ao refletir o comportamento acumulado da qualidade dos dados, o LWQS fornece uma perspetiva mais estável e robusta da fiabilidade do sistema, sendo especialmente relevante em contextos industriais sujeitos a variações operacionais frequentes.

A métrica é formalizada pela equação 3.3:

$$LWQS_j = w_a \left[\sum_{s \in S} w_s \left(\frac{\sum_{k \in K} f_k \, a_k^s}{\sum_{k \in K} f_k \, n_k} \right) \right] + w_c \left[\sum_{s \in S} w_s \left(\frac{\sum_{k \in K} f_k \, c_k^s}{\sum_{k \in K} f_k \, n_k} \right) \right] \tag{3.3}$$

No presente estudo, considerou-se uma janela de $m = 12$ blocos (correspondente a 1 hora) e foram definidos pesos $w_s = 0{,}019$ para os sensores $s = 1, ..., n_s - 1$, sendo o último sensor ajustado para $w_{n_s} = 1 - 0.019(n_s - 1)$. Este modelo permite que sensores críticos recebam maior importância no cálculo, adaptando-se às exigências específicas do sistema. O LWQS é calculado com base numa média ponderada de blocos anteriores, utilizando um fator de decaimento exponencial definido por $f(k) = \exp\left(-\frac{j-k-1}{\beta}\right)$ em que $\beta > 0$ controla a taxa de decaimento e determina o grau de influência dos blocos mais antigos. Esta estratégia permite ao sistema reagir mais rapidamente a alterações recentes na qualidade dos dados, respeitando simultaneamente a natureza volátil e muitas vezes imprevisível dos ambientes industriais monitorizados por sistemas IdC.

A utilização do LWQS possibilita não só acompanhar a trajetória da qualidade dos dados, como também detetar degradações ou melhorias subtis que poderiam passar despercebidas numa análise meramente instantânea. Ao conjugar granularidade temporal, ponderação por sensor e sensibilidade à evolução das métricas, o LWQS afirma-se como uma ferramenta essencial para a monitorização contínua e para uma gestão proativa da qualidade dos dados em sistemas industriais.

3.2.4 Análise dos Resultados

A Figura 3.9 apresenta os valores obtidos para os índices de qualidade calculados com base nas métricas apresentadas, ao longo do mês de abril de 2018. Observa-se que o QSD (parte inferior da figura) evidencia uma tendência geral de estabilidade, embora consiga identificar de forma eficaz picos positivos e negativos na qualidade dos blocos analisados. Destaca-se que, imediatamente antes das falhas do sistema (assinaladas com linhas vermelhas), tanto o WQS como o LWQS revelam uma queda, sendo que o QSD assume valores negativos. Este padrão sugere que o sistema estava a entrar numa fase de degradação, o que valida o potencial do índice como mecanismo de alerta precoce (4).

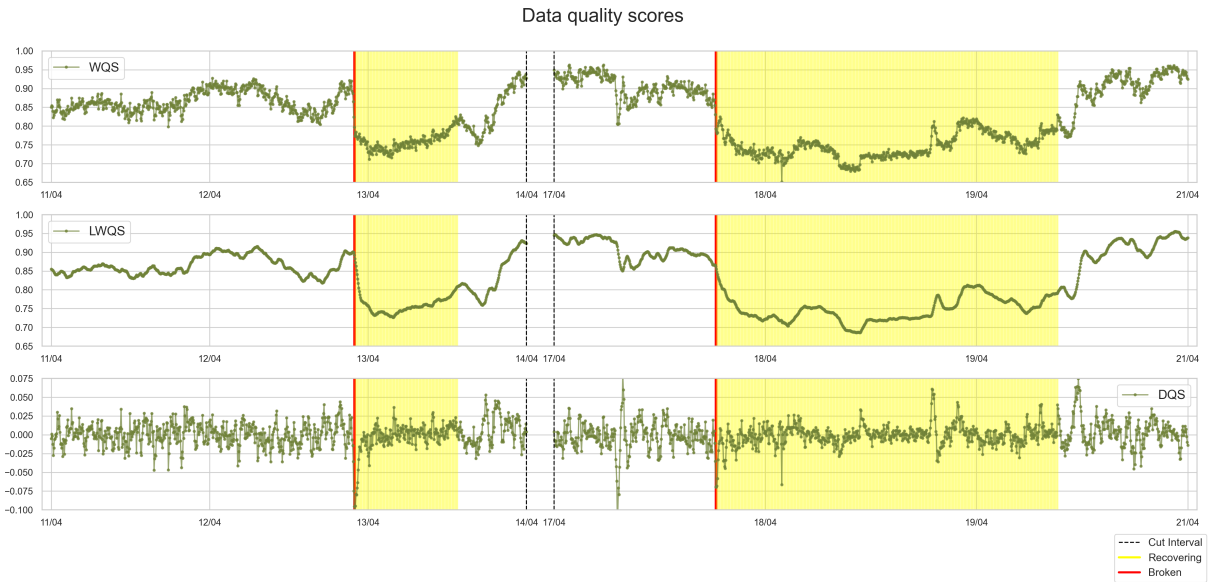


Figura 3.9: WQS (topo), LWQS (meio) e QSD (base). Fonte: (4)

De acordo com os autores (4), o aumento do parâmetro m, correspondente ao número de blocos considerados na janela histórica, pode potenciar a capacidade de deteção precoce de degradações na qualidade dos dados, permitindo uma identificação mais atempada de padrões de deterioração. As equações 3.2 e 3.3 viabilizam ainda a atribuição de pesos diferenciados por sensor, o que representa uma mais-valia em contextos onde se pretende refletir a criticidade operacional de determinados dispositivos. No presente estudo, uma vez que não foi disponibilizada informação específica sobre a relevância relativa dos sensores, optou-se por uma ponderação uniforme, conduzindo a resultados similares entre as diferentes formulações (4).

Os resultados obtidos demonstram que os índices de qualidade propostos constituem métricas flexíveis e adaptáveis à avaliação da qualidade dos dados em ambientes industriais baseados em IdC. Ao contrário de abordagens tradicionais baseadas em critérios fixos, o QSD apresenta-se como uma solução dinâmica, ajustável a diferentes contextos e necessidades de dados. A sua arquitetura modular permite ainda a integração de outras dimensões, como consistência ou atualidade, facilitando a adaptação a cenários IdC diversos, onde a importância relativa de cada dimensão pode variar substancialmente (4).

Com base nos resultados, conclui-se que os índices WQS, LWQS e QSD são capazes de integrar eficazmente múltiplas métricas de qualidade, fornecendo uma visão abrangente e coerente da qualidade dos dados (4). A sua aplicação permite às organizações detetar de

forma eficiente problemas de dados, apoiar a tomada de decisão em tempo real e sustentar estratégias de melhoria contínua nos processos de gestão de dados.

Com este segundo caso de estudo, foi possível validar, num cenário realista e com dados públicos, a aplicabilidade dos índices compostos WQS, LWQS e QSD no acompanhamento contínuo da qualidade dos dados. Através da integração de dimensões e da incorporação de perspetivas temporais, demonstrou-se a capacidade do modelo proposto em refletir o comportamento dinâmico da qualidade dos dados, reforçando a sua utilidade prática em ambientes industriais. Esta evolução metodológica, face ao primeiro estudo mais controlado, consolida os fundamentos teóricos apresentados e abre caminho para a aplicação da abordagem em contextos operacionais ainda mais complexos, como o que será explorado no caso de estudo seguinte.

3.3 Caso de Estudo 3 – Arquitetura de Monitorização da Qualidade de Dados em Tempo Real

A crescente digitalização dos processos industriais, impulsionada pelos princípios da I4.0, trouxe consigo a necessidade de monitorizar continuamente grandes volumes de dados provenientes de sensores em tempo real (5). Neste contexto, a qualidade dos dados assume um papel crítico, influenciando diretamente a fiabilidade dos diagnósticos, a eficiência operacional e a capacidade de antecipar falhas através de estratégias de manutenção preditiva. Este caso de estudo foca-se na aplicação e validação de uma arquitetura modular e leve para avaliação contínua da qualidade dos dados em ambientes industriais reais.

O objetivo principal consistiu em validar e integrar, num único *pipeline*, os conceitos, métricas e mecanismos de monitorização desenvolvidos nos trabalhos anteriores, assegurando a sua aplicabilidade em tempo real e em contexto de produção, e é baseado no artigo ***A Data Quality Pipeline for Industrial Environments: Architecture and Implementation*** de Peixoto et al. (5).

Este caso de estudo final corresponde a uma linha de produção automatizada equipada com motores elétricos responsáveis pelo funcionamento de sistemas transportadores. Um sistema transportador é um equipamento mecânico comum que transporta materiais de um local para outro (77). Estes sistemas são amplamente utilizados na indústria para movimentar materiais pesados ou volumosos entre diferentes etapas de produção (77). A tecnologia de transporte é utilizada em várias aplicações, incluindo passadeiras rolantes e escadas rolantes, bem como em muitas linhas de montagem de fabrico (77). Além disso, os transportadores são amplamente utilizados em sistemas automatizados de distribuição e armazenamento (77). Cada sistema de transporte monitorizado é acionado por um motor específico, identificado como $SE01$, $SE02$ ou $SE03$, e equipado com sensores de temperatura e vibração.

No presente estudo, a monitorização centrou-se exclusivamente nos sensores de temperatura, denominados $Temp1$, $Temp2$ e $Temp3$, por se tratarem de variáveis críticas para a deteção de sobreaquecimentos, identificação de padrões térmicos anómalos e suporte a estratégias de manutenção preditiva (5). Estes sensores foram configurados para recolher uma leitura por segundo, gerando um fluxo contínuo e de alta frequência de dados. Cada registo é estruturado no formato *JSON* e inclui quatro campos fundamentais: identificador do sensor, nome da variável, valor medido e *timestamp* da leitura, conforme ilustrado

na Tabela 3.3.

Tabela 3.3: Descrição das variáveis recolhidas num sistema transportador.

Campo	**Tipo**	**Descrição**
`código`	*string*	Identificador do sensor
`nome`	*string*	Nome da variável
`valor`	*float*	Valor medido
`data`	*timestamp*	Instante da medição

Um exemplo de mensagem recebida em tempo real:

```
{
"code": "Temp1",
"name": "Temp1",
"value": 42.5,
"date": "2025-05-23T09:07:36.627000"
}
```

Esta estrutura simples, mas expressiva, facilita a uniformização do processamento e permite a integração eficiente em arquiteturas de ingestão de dados em tempo real. A utilização de três sensores independentes em equipamentos distintos permite, adicionalmente, realizar análises comparativas e identificar padrões de comportamento entre componentes equivalentes do sistema.

A elevada cadência de aquisição torna possível observar o comportamento térmico dos motores com grande granularidade, mas também amplifica os desafios inerentes à qualidade dos dados. Entre os problemas mais frequentemente identificados neste cenário estão a presença de valores extremos ou anómalos, falhas pontuais na comunicação dos sensores, intervalos com ausência de dados e discrepâncias temporais entre registos consecutivos (5). Estas situações são comuns em ambientes industriais devido à complexidade dos sistemas físicos e às condições operacionais adversas, como ruído eletromagnético, vibrações excessivas ou limitações de conectividade (16).

A Figura 3.10 ilustra um excerto real dos dados recolhidos pelos sensores $Temp1$, $Temp2$ e $Temp3$, entre as 12h10 e as 18h00 do dia 24 de maio de 2025. No painel esquerdo observa-se a evolução temporal das leituras de temperatura, enquanto no painel direito é apresentada uma tabela com os registos brutos, incluindo o *timestamp*, o valor medido e o identificador do sensor correspondente.

Esta visualização conjunta permite não só a inspeção visual de tendências e variações, como também a identificação de anomalias, como picos abruptos, quedas súbitas de valor e lacunas temporais entre medições. Estas irregularidades podem indicar falhas de sensor, interrupções na transmissão dos dados, erros de calibração ou outras disfunções no sistema. A sua deteção precoce é fundamental para evitar decisões baseadas em dados incorretos ou desatualizados.

Neste contexto, a necessidade de mecanismos robustos para avaliação contínua da qualidade dos dados torna-se evidente. Sem tais mecanismos, as decisões baseadas em dados

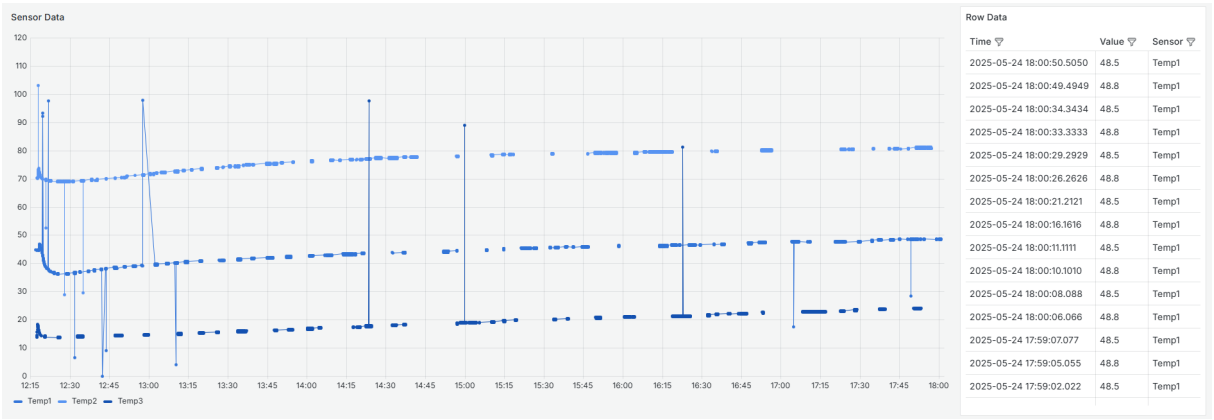


Figura 3.10: Dados recolhidos pelos 3 sensores de temperatura. Fonte: (5).

industriais podem tornar-se imprecisas ou até perigosas para os processos em causa (1). A arquitetura testada neste caso de estudo oferece uma resposta prática a esse desafio, permitindo não apenas visibilidade sobre a evolução da qualidade dos dados ao longo do tempo, mas também a obtenção de indicadores acionáveis que sustentam estratégias de melhoria contínua.

Assim, este caso de estudo representa a consolidação dos conceitos desenvolvidos ao longo deste trabalho, demonstrando que é possível integrar, de forma eficiente e escalável, todos os componentes de uma solução de monitorização da qualidade dos dados numa arquitetura leve, extensível e orientada para ambientes industriais em tempo real.

3.3.1 Arquitetura

A arquitetura desenvolvida foi concebida para dar resposta às exigências específicas dos ambientes industriais, caracterizados por grandes volumes de dados, elevada frequência de aquisição e heterogeneidade de fontes. Inspirada nos princípios modulares descritos por Oliveira et al. (69; 8), a solução proposta integra três serviços principais, ingestão, avaliação da qualidade e visualização, que operam de forma paralela sobre uma infraestrutura orientada a eventos (5).

A Figura 3.11 apresenta a arquitetura implementada. Esta estrutura recorre a tecnologias consolidadas e *open-source*, nomeadamente o *Apache Kafka*[5] para transporte de dados, o *InfluxDB*[6] para armazenamento e o *Grafana*[7] para visualização. Este conjunto permite não só garantir baixa latência e escalabilidade, como também assegurar a monitorização contínua da qualidade dos dados gerados em tempo real pelos sensores industriais.

O processo inicia-se com a leitura dos sensores industriais e a publicação direta dos dados brutos num *broker* central do *Kafka*, evitando a necessidade de pré-processamento intermédio. Neste caso, foi utilizado apenas um tópico (*sensors.temperature-vX*), dado que o sistema monitorizado incluía exclusivamente sensores de temperatura. Cada sensor funciona como uma fonte autónoma, e este modelo *source-to-broker* garante baixa latência, elevada escalabilidade e rastreabilidade desde o ponto de origem dos dados (5).

[5]Apache Kafka: `https://kafka.apache.org/`
[6]InfluxDB: `https://www.influxdata.com/`
[7]Grafana: `https://grafana.com/`

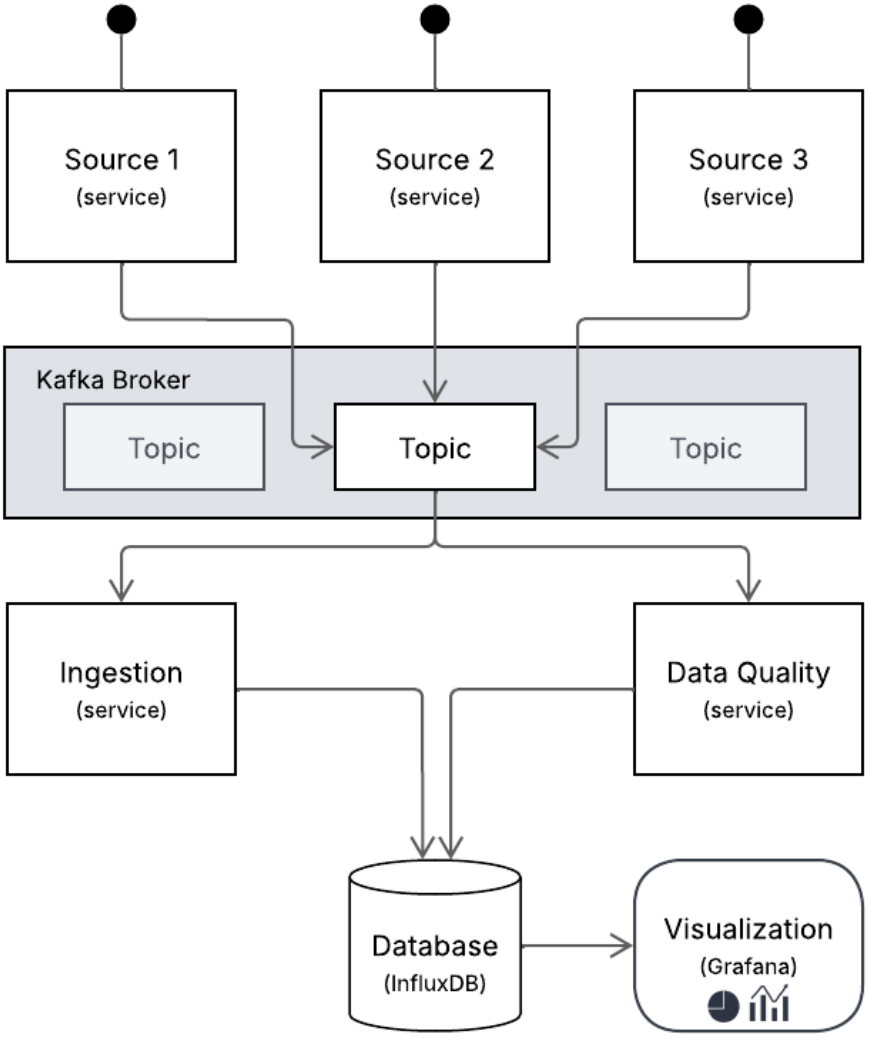


Figura 3.11: Arquitetura de monitorização da qualidade de dados. Fonte: (5).

O serviço de ingestão consome continuamente estas mensagens, interpreta os seus campos e armazena os dados no *InfluxDB*, uma base de dados temporal otimizada para cargas de trabalho intensivas e análises em tempo real. Ao contrário de abordagens anteriores que recorriam à validação com esquemas AVRO (8), esta implementação prioriza a aquisição em tempo real, adiando os mecanismos de validação para um processo paralelo (5). Apesar de não validar os dados à entrada, a integridade dos mesmos é assegurada indiretamente na etapa seguinte, através da deteção de anomalias, lacunas e inconsistências de formatação durante o *data profiling*. Os registos brutos são armazenados no `bucket: data`, garantindo persistência e acesso eficiente para análises subsequentes. Além disso, os dados são encaminhados para blocos temporais de um minuto, que funcionam como janelas de agregação para os processos de avaliação da qualidade.

Estes blocos temporais são processados por um segundo serviço, o módulo de avaliação da qualidade, responsável por executar rotinas contínuas de *data profiling* e cálculo de métricas sobre os registos recebidos. Este serviço opera em paralelo com a ingestão, consumindo o mesmo fluxo de mensagens do tópico *Kafka* e avaliando cada bloco de dados com base em janelas de tempo deslizantes de um minuto. Para cada intervalo temporal, é

calculado um conjunto abrangente de métricas específicas das principais DQD, acurácia, completude, consistência e atualidade.

As métricas são calculadas de forma incremental e agregadas em índices compostos que permitem uma visão consolidada e evolutiva do desempenho dos sensores. A avaliação da qualidade é realizada com base nos dados armazenados no `bucket: data`, sendo os resultados registados no `bucket: data quality`. Para efeitos de análise histórica, considerou-se o último dia de registos disponíveis no *InfluxDB*. Todos os resultados apresentados referem-se a dados recolhidos entre as 12h10 e as 15h30 do dia 24 de maio de 2025, um intervalo com elevada variabilidade, ideal para avaliação crítica da qualidade dos dados.

Cada dimensão foi operacionalizada através de métricas desenvolvidas nos casos de estudo anteriores, concebidas para refletir de forma objetiva o grau de conformidade dos dados em relação aos requisitos de fiabilidade, integridade estrutural, coerência e oportunidade. A seguir, apresentam-se de forma concisa as métricas adotadas para cada dimensão, bem como as adaptações introduzidas na presente implementação.

No caso da acurácia, a métrica (Eq. 2.2) baseia-se na normalização dos valores registados por cada sensor, utilizando os limites mínimo e máximo do respetivo histórico, onde os resultados fora do intervalo $[0, 1]$ são classificados como anómalos e destacados visualmente nos *dashboards*. Contrariamente aos estudos anteriores, em que os limites eram definidos com base em percentis fixos (10.º e 90.º), e posteriormente através da regra empírica dos 68–95–99.7%, neste estudo recorreu-se a uma diferente abordagem: o filtro de *Hampel*. Esta técnica, baseada na mediana e no Desvio Absoluto da Mediana (MAD), permite mitigar a influência de *outliers* e aumentar a fiabilidade da normalização. O intervalo de aceitação é definido como $I = [\text{mediana} - 3 \times \text{MAD}, \text{mediana} + 3 \times \text{MAD}]$.

Os valores mínimo e máximo utilizados na equação da acurácia são definidos dinamicamente com base nos limites identificados pelo filtro de *Hampel*. Este método utiliza um filtro estatístico para garantir que valores extremos não distorcem o processo de normalização (78). A escolha do limiar de três desvios MAD resulta de um equilíbrio entre sensibilidade à deteção de anomalias e robustez em ambientes industriais ruidosos e instáveis (79). Esta abordagem revelou-se especialmente adequada para dados de sensores, frequentemente afetados por ruído e comportamentos esporádicos.

A Figura 3.12 apresenta uma comparação visual detalhada entre três abordagens distintas aplicadas aos três sensores de temperatura: percentis (10.º e 90.º), quantis extremos (0.00135 e 0.99865) e o filtro de *Hampel*. Esta estrutura de comparação foi inspirada na metodologia explorada por Peixoto et al. (5), permitindo avaliar de forma sistemática o impacto de diferentes estratégias de limitação sobre a sensibilidade da métrica de acurácia.

Observa-se que a estratégia baseada no filtro de *Hampel* permite uma separação mais clara entre valores dentro do intervalo esperado (pontos verdes) e valores considerados anómalos (pontos amarelos e vermelhos). Esta abordagem revelou um desempenho equilibrado, ao conseguir detetar desvios relevantes sem amplificar flutuações naturais dos sensores. Em contraste, a abordagem dos quantis extremos, conduz a uma distribuição mais compacta dos valores (entre 0.35 e 0.55, por exemplo, no $Temp1$), atenuando variações mas também ocultando possíveis desvios significativos. Já os percentis evidenciam menor robustez em momentos de variação rápida, com maior suscetibilidade à distorção causada por *outliers*. Esta análise sustenta a escolha do filtro de *Hampel* como a solução mais adequada para ambientes industriais voláteis. A sua capacidade de adaptação dinâmica aos dados

Figura 3.12: Comparação visual da métrica de acurácia para os sensores $Temp1$, $Temp2$ e $Temp3$, utilizando três abordagens distintas. Figura inspirada em (5).

históricos de cada sensor, aliada à resiliência contra valores extremos e ruído, tornam esta abordagem preferível face aos outros métodos. Esta comparação não só valida a decisão metodológica adotada, como também fornece um quadro de referência visual útil para aplicações futuras de monitorização da acurácia.

As Figuras 3.13–3.15 apresentam a evolução da acurácia ao longo do tempo para os três sensores de temperatura durante o intervalo de análise. Os pontos coloridos representam o estado de cada bloco temporal: verde (dentro do intervalo esperado), amarelo (pequenos desvios) e vermelho (anomalias severas). O sensor $Temp1$ apresentou flutuações acentuadas no início do período, com vários blocos assinalados a amarelo, em concordância com as oscilações visíveis na Figura 3.10. Com o tempo, os valores estabilizam, refletindo um comportamento térmico mais consistente.

Figura 3.13: Evolução da métrica de acurácia para o sensor $Temp1$ ao longo do tempo. Fonte: (5)

No caso do sensor $Temp2$, os valores de acurácia começam mais baixos, sugerindo leituras inicialmente subestimadas, mas estabilizam progressivamente após as 13h00, sem registos classificados como anómalos. Já o sensor $Temp3$ evidenciou instabilidade em dois momentos (cerca das 14h20 e 15h00), com pequenas oscilações associadas a picos de temperatura,

também observáveis na Figura 3.10. No entanto, após esses episódios, o sensor manteve um desempenho estável.

Figura 3.14: Evolução da métrica de acurácia para o sensor $Temp2$ ao longo do tempo. Fonte: (5)

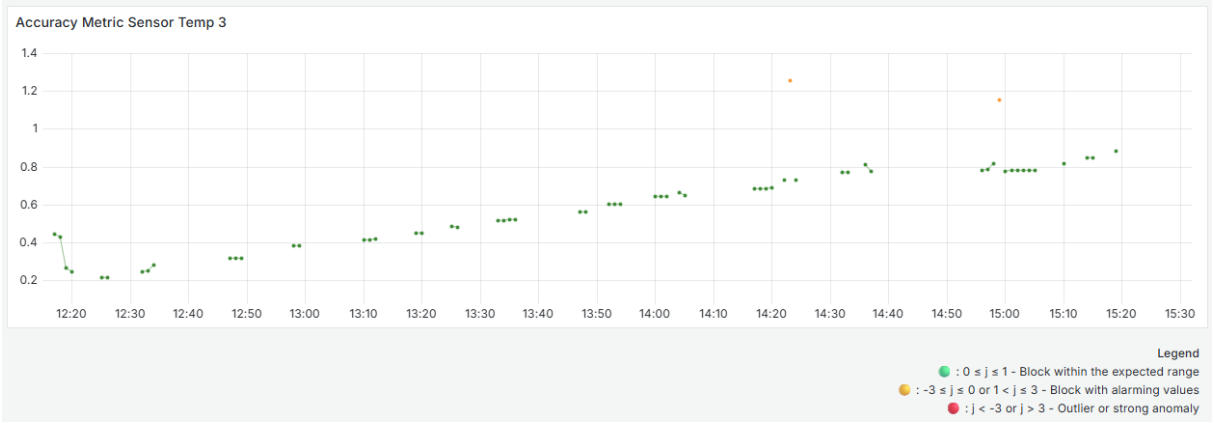


Figura 3.15: Evolução da métrica de acurácia para o sensor $Temp3$ ao longo do tempo. Fonte: (5)

Importa salientar que, ao longo da janela analisada, não foram detetados blocos com pontos vermelhos, o que indica que, apesar das variações e de alguns desvios localizados, nenhum dos blocos foi considerado gravemente comprometido do ponto de vista da acurácia.

Relativamente às restantes dimensões da qualidade, completude (métrica 2.3 e 2.4), consistência (métrica 2.5) e atualidade (métrica 2.9), importa referir que não foram introduzidas alterações nas fórmulas de cálculo nem nos critérios de avaliação face aos casos de estudo anteriores. A sua aplicação neste terceiro caso seguiu exatamente os mesmos princípios metodológicos, com adaptação apenas ao novo fluxo de dados em tempo real.

No caso da completude, foram utilizadas duas métricas complementares. A primeira avalia a completude ao nível do conteúdo (métrica 2.3), verificando a presença de todos os campos obrigatórios nos registos e a inexistência de valores nulos ou vazios. A segunda métrica calcula a completude temporal (métrica 2.4), medindo a proporção de registos efetivamente recebidos face ao número esperado por intervalo de tempo, com base na cadência de aquisição configurada (60 registos por minuto por sensor). Esta distinção

permite avaliar separadamente a integridade estrutural dos dados e a regularidade da sua transmissão.

A consistência foi calculada com base nas correlações entre os dados históricos dos sensores (métrica 2.5), utilizando o coeficiente de Pearson para identificar relações fortes (acima de 0.8). Estas relações são registadas como regras, que são depois comparadas com os dados atuais em cada bloco de tempo. Novas regras podem ser adicionadas automaticamente sempre que se detetem correlações consistentes, permitindo que o sistema se adapte ao comportamento dos sensores.

A atualidade considerou o tempo decorrido entre a geração e a receção de cada registo, bem como a volatilidade dos dados, definida neste estudo como uma janela de dois minutos. Esta parametrização permite adaptar a métrica a diferentes graus de sensibilidade temporal, tornando-a adequada tanto para contextos industriais em tempo real como para cenários com maior tolerância à latência. Esta abordagem permite avaliar não apenas a frescura da informação, mas também a sua utilidade no momento em que é analisada no contexto operacional. A métrica de atualidade (Eq. 2.9) produz valores normalizados entre 0 e 1, sendo 1 o valor ideal.

As figuras ilustrativas da evolução destas métricas ao longo do tempo, bem como os respetivos exemplos de blocos classificados, encontram-se disponíveis no artigo (5), onde é possível observar a estabilidade geral do sistema e a ocorrência pontual de desvios temporais ou perdas de completude associadas a falhas de comunicação.

A monitorização contínua da qualidade dos dados em ambientes industriais requer não apenas a avaliação isolada de cada métrica, mas também a sua integração num índice sintético que permita interpretar rapidamente o estado geral do sistema. Com esse propósito, foram utilizados os três indicadores já antes referidos: o WQS, o LWQS e o QSD. Estes índices foram inicialmente propostos por Costa e Silva et al. (76), posteriormente desenvolvidos e adaptados em (4), e agora, em (5), voltam a ser ajustados na sua implementação, de forma a refletir melhor a realidade dinâmica e heterogénea do sistema em estudo.

O WQS_j (Eq. refeq:wsq) calcula a qualidade dos dados no bloco j, agregando os resultados de acurácia e completude de todos os sensores. Já o LWQS_j (Eq. 3.3) reflete a qualidade histórica acumulada dos m blocos anteriores, ponderando as observações mais recentes com maior peso, através de uma função de decaimento exponencial. A diferença entre estes dois valores dá origem ao QSD_j (Eq. 3.1), que mede a variação recente na qualidade dos dados, permitindo detetar melhorias ou degradações ao longo do tempo.

Uma das principais inovações nesta versão foi a introdução de pesos específicos por sensor e por dimensão. Ao contrário da abordagem original, em que se utilizavam apenas pesos globais para acurácia (w_a) e completude (w_c), aqui definem-se pesos individuais para cada sensor: w_a^s e w_c^s, com a condição de que $\sum w_d^s = 1$ para cada dimensão $d \in a, c$. Esta separação permite refletir a criticidade operacional de cada sensor e a relevância relativa das dimensões, adaptando o índice ao contexto do sistema monitorizado.

Além disso, a fórmula de cálculo da completude foi ajustada, substituindo-se a métrica anteriormente utilizada pela métrica 2.4. Em vez da razão c_j^s/n_j^s (número de registos completos sobre os recebidos), passou a ser usada c_j^s/e_j^s, onde e_j^s representa o número esperado de registos do sensor s no bloco j. Esta alteração fornece uma estimativa mais fiável da perda de dados, ao comparar diretamente o que foi efetivamente recebido com o

que era esperado.

Para o componente histórico (LWQS), consideraram-se $m = 10$ blocos (correspondendo a 10 minutos), com uma função de decaimento exponencial, sendo $\beta = 5$ o valor adotado. Este parâmetro foi definido com base na inspeção visual da curva de pesos (apresentada no artigo (5)), por oferecer um bom compromisso entre sensibilidade a alterações recentes e preservação de contexto histórico.

Os pesos globais utilizados foram $w_a = 0.7$ para acurácia e $w_c = 0.3$ para completude, refletindo a maior influência da acurácia na qualidade global dos dados. Os pesos por sensor na componente de acurácia foram definidos como uniformes $(1/n_s)$ entre os sensores que efetivamente enviaram dados no bloco analisado, garantindo equilíbrio na ausência de prioridades explícitas. Já na completude, e de forma a penalizar os sensores que não enviam dados, considerou-se $n_s = 3$, dado que o sistema monitorizado incluía três sensores.

As fórmulas finais utilizadas foram:

$$QSD_j = \begin{cases} WQS_j - LWQS_j & \text{if } j > 1 \\ 1 & \text{if } j = 1 \end{cases} \tag{3.4}$$

$$WQS_j = w_a \left(\sum_{s \in S} w_a^s \, \frac{a_j^s}{n_j^s} \right) + w_c \left(\sum_{s \in S} w_c^s \, \frac{c_j^s}{e_j^s} \right) \tag{3.5}$$

$$LWQS_j = w_a \left[\sum_{s \in S} w_a^s \left(\frac{\sum_{k \in K} f_k \, a_k^s}{\sum_{k \in K} f_k \, n_k^s} \right) \right] + w_c \left[\sum_{s \in S} w_c^s \left(\frac{\sum_{k \in K} f_k \, c_k^s}{\sum_{k \in K} f_k \, e_k^s} \right) \right] \tag{3.6}$$

Estes três índices, WQS, LWQS e QSD, foram calculados em tempo real, bloco a bloco, permitindo acompanhar tendências, identificar anomalias e priorizar intervenções. Os resultados obtidos encontram-se detalhados e ilustrados no artigo de Peixoto et al. (5), com visualizações temporais que permitem analisar a evolução da qualidade dos dados ao longo do período monitorizado.

O WQS revela flutuações significativas de curto prazo na qualidade dos dados. O valor deste índice nunca ultrapassa 0.8, refletindo o impacto da baixa completude temporal registada ao longo do período. Em contraste, o LWQS, devido ao seu caráter histórico, apresenta uma curva mais suave, respondendo de forma mais gradual a degradações pontuais da qualidade. Ainda assim, consegue refletir o decréscimo geral causado pelas anomalias recentes, embora amortecido pela influência de blocos anteriores com melhor desempenho. O QSD, definido como a diferença entre o WQS e o LWQS, destaca-se como um indicador sensível à variação da qualidade em tempo real. Valores positivos indicam melhorias face ao histórico recente, enquanto valores negativos sinalizam deteriorações. A predominância de pontos amarelos ao longo do gráfico sugere uma estabilidade relativa, com poucas variações significativas no desempenho global do sistema.

A utilização de pesos específicos por sensor e por dimensão permite que estes índices reflitam com maior precisão a relevância operacional de cada componente do sistema. Esta abordagem torna os indicadores especialmente adequados para ambientes industriais heterogéneos e sujeitos a flutuações imprevisíveis, como os baseados em IdC, onde a

elevada variabilidade e a necessidade de resposta rápida exigem mecanismos capazes de captar tanto o estado atual como a evolução da qualidade dos dados ao longo do tempo.

O último componente da arquitetura é o módulo de visualização, implementado com recurso à plataforma *Grafana*, que atua como interface principal para os utilizadores. Os *dashboards* desenvolvidos permitem acompanhar em tempo real tanto os valores brutos recolhidos pelos sensores como os indicadores de qualidade, facilitando a deteção precoce de anomalias e a tomada de decisão informada. Os dados exibidos provêm diretamente do *InfluxDB*, garantindo consultas rápidas e eficazes sobre séries temporais. Esta integração permite que operadores e engenheiros de processo visualizem o estado da produção com base em dados fiáveis e atualizados, sem necessidade de intervenção técnica adicional.

Aqui foram desenvolvidos dois *dashboards* interativos em *Grafana* para suporte à análise em tempo real, conforme ilustrado nas Figuras 3.16 e 3.17 e desenvolvido em (5).

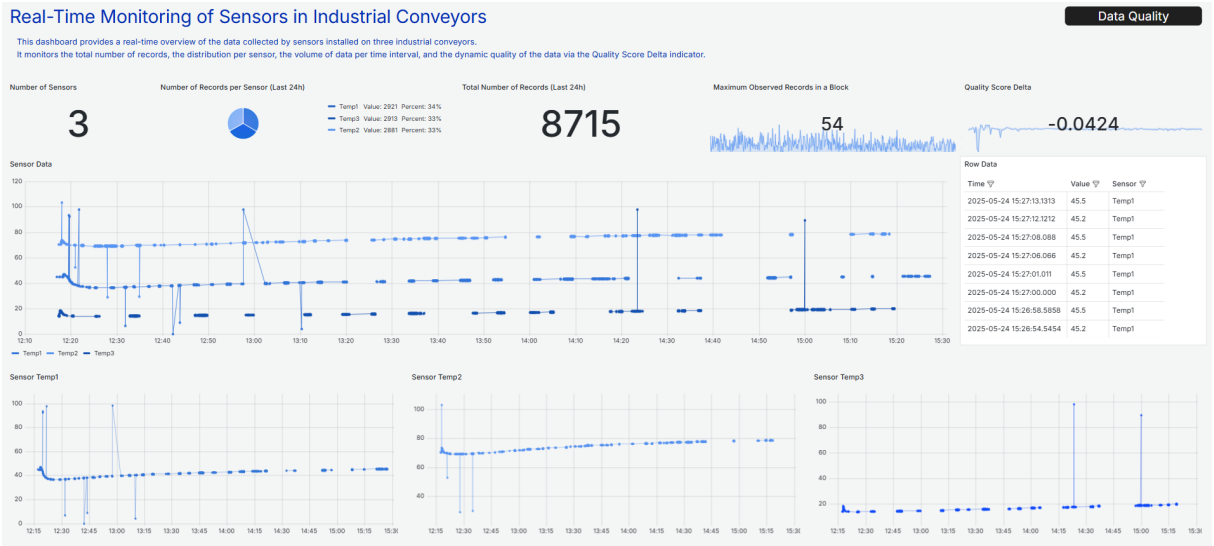


Figura 3.16: *Dashboard* operacional de dados brutos. Fonte: (5)

A Figura 3.16 apresenta uma visão geral dos dados brutos armazenados no `bucket: data`, incluindo informação agregada sobre o número total de registos recebidos nas últimas 24 horas, contagem por sensor, uma tabela com os registos brutos, gráficos globais e visões individuais para cada sensor.

Em contraste, a Figura 3.17 foca-se exclusivamente na monitorização da qualidade dos dados. Os três painéis superiores exibem os principais índices de qualidade (WQS, LWQS, QSD), seguidos de um gráfico consolidado com a evolução temporal das métricas. A tabela central destaca avisos e desvios face aos limiares definidos, sendo ainda disponibilizados gráficos individuais por métrica. Esta separação entre os dados brutos e os indicadores de qualidade facilita a supervisão do sistema, permitindo uma interpretação mais clara e uma identificação mais célere de problemas.

Estes *dashboards* foram concebidos para apoiar a tomada de decisão operacional, oferecendo filtros temporais, seleção de sensores específicos e indicadores visuais com codificação por cores. A presença de limiares de alerta e marcadores gráficos permite a deteção imediata de anomalias, fluxos de dados incompletos ou falhas nos sensores. Este ambiente interativo garante que os problemas de qualidade são rapidamente identificados e resolvidos, reduzindo o tempo de diagnóstico e promovendo uma manutenção mais proativa do

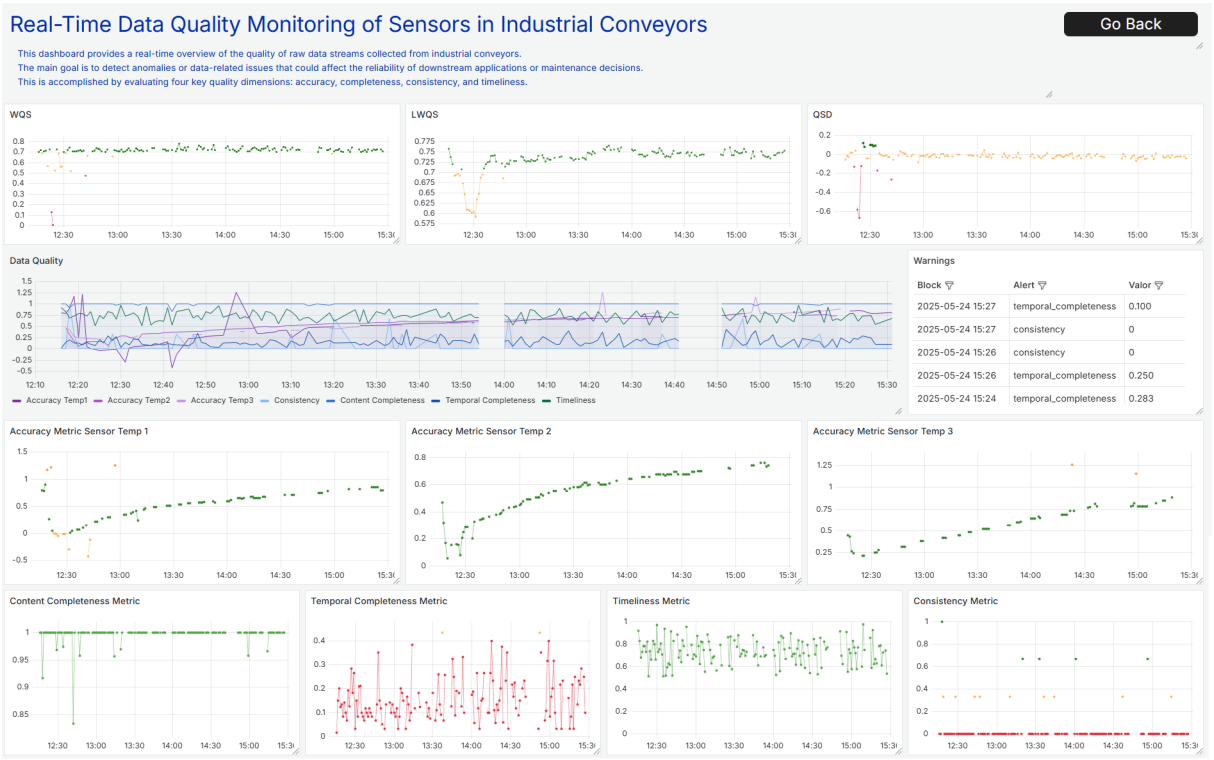


Figura 3.17: *Dashboard* de monitorização da qualidade. Fonte: (5)

sistema.

A arquitetura proposta, composta por serviços modulares para ingestão, avaliação da qualidade e visualização, assenta num modelo orientado a eventos com base em *Apache Kafka*, o que assegura baixa latência, elevada escalabilidade e tolerância a falhas. A separação entre os processos de aquisição e validação permite a evolução independente de cada componente, aumentando a flexibilidade do sistema. Esta abordagem torna a solução especialmente adequada a ambientes industriais dinâmicos, onde a heterogeneidade das fontes de dados e a exigência de resposta em tempo real representam desafios significativos à fiabilidade e utilidade dos dados recolhidos.

3.3.2 Análise dos Resultados

Durante os testes realizados, a arquitetura demonstrou capacidade para suportar a ingestão contínua de dados com latência mínima. As métricas aplicadas evidenciaram flutuações relevantes, com quebras na completude temporal sempre que os sensores deixavam de enviar dados, reduções na consistência quando as correlações esperadas entre sensores não eram verificadas, e deteção de *outliers* em sensores específicos, indiciando potenciais falhas ou leituras imprecisas. Este nível de granularidade permite um diagnóstico detalhado da qualidade dos dados em ambiente industrial, reforçando o valor prático da solução. A separação entre os serviços de ingestão e de avaliação da qualidade, e a sua execução em paralelo, garantem robustez e escalabilidade (5).

A leveza da arquitetura e a sua flexibilidade tornam-na particularmente adequada para ambientes industriais com recursos computacionais limitados (5). A solução não depende de frameworks declarativas externas, integrando os mecanismos de validação diretamente

no serviço de qualidade através do *data profiling*. Esta abordagem é extensível a outros tipos de sensores ou variáveis industriais, como vibração ou pressão, sem exigir alterações estruturais significativas. Naturalmente, poderão ser necessárias adaptações específicas consoante o setor de aplicação, por exemplo, sensores de torque e vibração podem ser prioritários no setor automóvel, enquanto a completude e conformidade de parâmetros por lote podem assumir maior relevância na indústria alimentar (5). A estrutura modular do *pipeline* facilita estas adaptações com esforço mínimo.

Apesar das suas vantagens práticas, a arquitetura apresenta algumas limitações. A dependência de janelas temporais de processamento pode introduzir atrasos ligeiros no cálculo das métricas, limitando a aplicabilidade em contextos que exijam latência ultra-baixa. Embora o sistema tenha demonstrado desempenho consistente durante os testes realizados, poderão ser necessárias otimizações adicionais para ambientes de carga extrema, como fluxos com milhares de sensores em simultâneo. Nesses casos, recomenda-se ajustar o particionamento de *Kafka*, os grupos de consumidores e as estratégias de indexação no *InfluxDB*, de forma a evitar constrangimentos no processo de ingestão ou armazenamento de dados. Adicionalmente, a solução ainda não foi validada em cenários com transições de estado da máquina ou contextos operacionais multimodais, onde o comportamento dos sensores pode variar significativamente consoante o modo de operação.

A arquitetura foi implementada com ferramentas consolidadas e de código aberto, *Apache Kafka*, *InfluxDB* e *Grafana*, garantindo desempenho satisfatório, robustez e visualização intuitiva. A modularidade e leveza da solução tornam-na adaptável a diferentes contextos industriais, mesmo perante restrições operacionais ou computacionais (5). Esta flexibilidade é fundamental para apoiar a evolução dos sistemas de monitorização no contexto da I4.0, onde a integração e escalabilidade são requisitos-chave. A sua aplicabilidade prática é clara em cenários onde a qualidade da informação tem impacto direto na eficiência operacional, na fiabilidade dos processos e na implementação de estratégias de manutenção preditiva. A solução desenvolvida constitui uma base sólida para análises mais avançadas baseadas em dados, apoiando uma transição mais segura e informada para a digitalização industrial.

Este caso de estudo demonstrou a aplicabilidade de uma arquitetura modular e leve para a monitorização contínua da qualidade dos dados num ambiente industrial real. A integração das métricas de acurácia, completude, consistência e atualidade num *pipeline* contínuo permitiu obter uma visão abrangente da qualidade dos dados, facilitando a deteção de falhas de forma automática e em tempo real. A flexibilidade e adaptabilidade da solução tornam-na adequada a diferentes contextos industriais e cenários operacionais.

Os resultados confirmam o potencial da abordagem proposta para apoiar a transformação digital das organizações, oferecendo uma base fiável para análises avançadas, otimização de processos e tomada de decisão baseada em dados. Esta implementação contribui para colmatar a distância entre a teoria e a prática no domínio da qualidade de dados no contexto da I4.0 (5).

Capítulo 4

Discussão dos resultados

Este capítulo apresenta uma análise crítica dos resultados obtidos nos três casos de estudo desenvolvidos ao longo desta dissertação. Cada caso foi concebido para explorar diferentes perspetivas da qualidade dos dados no contexto da I4.0, desde a aplicação de métricas individuais até à construção de um *pipeline* completo de monitorização em tempo real. Assim, esta discussão procura identificar padrões, pontos fortes e limitações comuns, refletindo sobre a aplicabilidade prática das abordagens propostas e a sua relevância para o avanço da digitalização na I4.0.

4.1 Evolução da Complexidade entre os Estudos

Os três casos de estudo apresentados nesta dissertação foram concebidos com níveis crescentes de complexidade, tanto em termos de contexto de aplicação como da sofisticação das abordagens metodológicas utilizadas. Esta progressão permitiu testar diferentes estratégias de avaliação da qualidade dos dados, desde a aplicação isolada de métricas até à integração contínua e automatizada num *pipeline* de monitorização em tempo real.

O Caso de Estudo 1 teve como objetivo avaliar a aplicabilidade de quatro métricas fundamentais, acurácia, completude, consistência e atualidade, num cenário controlado com dados simulados, possibilitando a verificação do comportamento das métricas em situações previamente conhecidas e a sua comparação com técnicas clássicas de deteção de anomalias, como o STL e o LSTM. Este estudo funcionou como uma base experimental, permitindo validar os conceitos fundamentais e explorar os limites de sensibilidade e robustez das métricas aplicadas.

O Caso de Estudo 2 introduziu uma maior complexidade, com a aplicação das métricas a um sistema de bombagem de água real, caracterizado por um volume elevado de sensores, dados e por falhas técnicas. Para além das métricas individuais, este estudo introduziu o conceito de índices compostos de qualidade, WQS, LWQS e QSD, capazes de agregar múltiplas dimensões num único valor interpretável. Esta abordagem permitiu uma visão mais holística da qualidade dos dados e uma comparação facilitada entre sensores, períodos e configurações.

Por fim, o Caso de Estudo 3 representa o culminar da abordagem, com a implementação de uma arquitetura modular e leve para a monitorização contínua da qualidade em ambiente industrial real. Este estudo integrou as métricas validadas nos casos anteriores num

pipeline em tempo real, utilizando tecnologias como *Apache Kafka*, *InfluxDB* e *Grafana*. A solução foi validada com dados operacionais e demonstrou aplicabilidade prática em contexto produtivo.

Esta evolução metodológica e tecnológica reflete o amadurecimento da abordagem proposta, desde a experimentação em ambiente simulado até à aplicação em sistemas reais, com impacto direto na operação e na tomada de decisão.

4.2 Comparação entre Métricas Aplicadas

Ao longo dos três casos de estudo, as DQD selecionadas, acurácia, completude, consistência e atualidade, foram aplicadas em contextos distintos, permitindo observar o seu comportamento e relevância em diferentes cenários industriais.

A métrica de acurácia foi transversal aos três estudos, com adaptações metodológicas e evoluções consoante o contexto. No Caso de Estudo 1, foi aplicada com base em limites históricos definidos por percentis e, posteriormente, por regras empíricas. Já no Caso de Estudo 3, foi introduzido o filtro de *Hampel*, mais robusto à presença de *outliers*, reforçando a fiabilidade dos resultados em ambiente real. Esta evolução metodológica demonstra a importância de adaptar os critérios de normalização consoante o grau de ruído e a variabilidade dos dados.

A completude teve impacto significativo em todos os casos de estudo, especialmente no diagnóstico de falhas de comunicação. No sistema de bombagem, a sua aplicação permitiu quantificar períodos de ausência de dados e associá-los a falhas técnicas previamente reportadas. No sistema real do Caso de Estudo 3, as quebras de completude revelaram-se frequentes e intermitentes, destacando a importância desta métrica em ambientes com infraestrutura envelhecida ou limitada.

A consistência revelou-se útil nos três casos de estudo, ao permitir identificar discrepâncias entre sensores cuja correlação era esperada. No entanto, é importante salientar que as regras foram sempre definidas com base nas correlações observadas nos dados, o que pode não ser suficiente. Para uma avaliação mais robusta, é recomendável envolver especialistas do domínio que possam definir regras de consistência com base no conhecimento dos processos operacionais e validar as correlações detetadas automaticamente.

Por fim, a atualidade foi particularmente relevante no Caso de Estudo 3, onde atrasos na chegada de dados podiam comprometer a capacidade de resposta em tempo real. Nos Casos de Estudo 1 e 2 esta dimensão teve uma aplicabilidade mais limitada, mas serviu como base para a definição de janelas de análise e validação da periodicidade dos registos.

Esta análise comparativa demonstra que, embora as métricas sejam teoricamente independentes, a sua utilidade prática depende fortemente das características do sistema monitorizado e dos objetivos operacionais definidos.

4.3 Índices de Qualidade

A introdução de índices compostos no Caso de Estudo 2, nomeadamente o WQS, o LWQS e o QSD, representou um avanço significativo na capacidade de análise e síntese da qualidade dos dados em contextos da I4.0. Estes índices permitiram integrar múltiplas métricas

individuais (neste caso, acurácia e completude) num único valor agregado, facilitando a interpretação dos resultados e a comparação entre sensores, períodos ou configurações operacionais.

O WQS forneceu uma avaliação instantânea da qualidade dos dados num determinado bloco temporal, refletindo a fiabilidade dos registos com base na sua acurácia e completude, ponderadas por pesos atribuídos a cada métrica e sensor. O LWQS introduziu uma perspetiva histórica, agregando os dados dos blocos anteriores com base numa função de decaimento exponencial, o que permitiu captar tendências e suavizar variações pontuais. Já o QSD, definido como a diferença entre o WQS e o LWQS, destacou-se como um indicador sensível à variação da qualidade dos dados ao longo do tempo, funcionando como um mecanismo de alerta precoce para eventuais degradações.

A aplicação destes índices no Caso de Estudo 2 validou a utilidade destes índices para monitorização contínua, permitindo a identificação de padrões de falha antes da sua ocorrência efetiva, como demonstrado pelas variações do QSD imediatamente antes de eventos críticos.

No Caso de Estudo 3, embora os índices WQS, LWQS e QSD não tenham sido reimplementados diretamente, foram alvo de pequenas adaptações, como a personalização dos pesos por sensor e a alteração da métrica utilizada para a completude. A arquitetura desenvolvida foi concebida para os suportar, refletindo uma evolução conceptual e técnica. As métricas foram integradas num *pipeline* em tempo real com base em *data profiling*, permitindo que os valores de qualidade fossem calculados continuamente e visualizados em *dashboards*. Este avanço abre caminho para que os índices compostos sejam recalculados de forma dinâmica, com possibilidade de incorporar outras dimensões, como consistência e atualidade, adaptando-se aos requisitos de sistemas industriais reais.

Assim, a evolução do Caso de Estudo 2 para o 3 demonstra a transição de uma abordagem baseada em blocos e análise retrospetiva para uma arquitetura orientada ao tempo real, escalável e preparada para suportar a integração de índices compostos e mecanismos de alerta dinâmicos.

A aplicação destes índices evidenciou várias vantagens práticas. Em primeiro lugar, permitiu uma rápida identificação de áreas problemáticas, sem a necessidade de análise métrica por métrica. Em segundo, revelou-se eficaz na comunicação de resultados a perfis menos técnicos, oferecendo uma visualização mais intuitiva da qualidade global dos dados. Além disso, os índices podem servir como base para a criação de alertas automáticos nos *dashboards*, aumentando a reatividade do sistema.

Contudo, estes índices também apresentaram limitações. A sua interpretação depende fortemente da qualidade das métricas subjacentes e da escolha dos pesos atribuídos, o que pode introduzir subjetividade no processo. Além disso, os índices não substituem a análise detalhada, sendo recomendável o seu uso em conjunto com visualizações complementares e diagnósticos mais específicos.

No geral, a introdução destes índices demonstrou ser uma ferramenta útil para consolidar a avaliação da qualidade dos dados, reforçando a capacidade analítica e a utilidade operacional das soluções desenvolvidas.

4.4 Impacto Prático, Limitações e Recomendações

Os resultados obtidos ao longo dos três casos de estudo demonstram o potencial das abordagens propostas para apoiar a monitorização e gestão da qualidade dos dados em ambientes industriais. A aplicação prática das métricas e índices de qualidade possibilitou a deteção de anomalias, falhas de comunicação e comportamentos irregulares de sensores, com impacto direto na eficiência operacional e na fiabilidade dos sistemas monitorizados. A visualização dos resultados através de *dashboards* contribuiu para uma interpretação mais acessível e para uma maior agilidade na resposta a problemas.

Além disso, a evolução das soluções, desde a avaliação de métricas individuais até à integração num *pipeline* contínuo, reflete uma clara progressão rumo a sistemas mais escaláveis, modulares e adaptáveis à realidade da I4.0. O desenvolvimento de índices compostos e a possibilidade de personalização por sensor e por métrica representam avanços significativos para a contextualização da qualidade dos dados e a adaptação a diferentes domínios industriais.

No entanto, foram também identificadas algumas limitações. Em primeiro lugar, a definição de regras de consistência baseadas exclusivamente em correlações extraídas dos dados pode carecer de validação por especialistas do domínio, o que compromete a fiabilidade dos resultados. Em segundo, os métodos de avaliação dependem da granularidade e qualidade da própria infraestrutura de recolha de dados, o que pode limitar a sua aplicação em sistemas mais antigos ou instáveis. Além disso, os pesos atribuídos nos índices compostos são sensíveis a critérios subjetivos, podendo enviesar os resultados caso não sejam devidamente fundamentados.

Adicionalmente, os testes realizados incidiram sobre um conjunto relativamente restrito de sensores, em particular, três sensores com elevada disponibilidade de dados, o que permitiu observar padrões de comportamento e testar o *pipeline* em condições favoráveis. No entanto, permanece incerta a forma como a arquitetura reagirá à escalabilidade horizontal, nomeadamente quando o número de sensores monitorizados aumenta significativamente e o volume de dados cresce de forma descontrolada. Esta incerteza levanta questões sobre a estabilidade, latência e consumo de recursos da solução em ambientes de produção mais exigentes.

De forma global, o *pipeline* proposto nesta dissertação demonstra ser não apenas viável tecnicamente, mas também relevante do ponto de vista operacional. A sua modularidade e adaptabilidade tornam-no aplicável a diferentes setores industriais, permitindo reforçar a confiança nos dados utilizados em processos críticos. Ao integrar a avaliação da qualidade dos dados como componente ativa de sistemas de monitorização, abre-se espaço para estratégias mais eficazes de manutenção preditiva, controlo de qualidade e otimização de recursos, promovendo uma transição mais segura e sustentada para a digitalização no contexto da I4.0.

Capítulo 5

Conclusões

Esta dissertação teve como principal objetivo investigar, aplicar e validar estratégias de avaliação da qualidade dos dados em ambientes industriais, no contexto da I4.0. Tendo em conta a crescente digitalização das operações industriais e a centralidade da informação na tomada de decisão, é imprescindível garantir que os dados recolhidos, armazenados e utilizados têm um nível de qualidade adequado aos objetivos operacionais e analíticos. Assim, com base na definição das dimensões mais relevantes da qualidade dos dados, acurácia, completude, consistência e atualidade, foram desenvolvidas métricas específicas, índices compostos e uma arquitetura modular de monitorização contínua, capaz de quantificar e interpretar, de forma automatizada, a fiabilidade dos dados provenientes de sensores industriais.

Ao longo dos três casos de estudo apresentados, foi possível validar a aplicabilidade das abordagens propostas em diferentes cenários, representando uma evolução metodológica gradual. O primeiro estudo, conduzido com dados simulados, teve como finalidade validar individualmente as métricas de qualidade e explorar a sua sensibilidade a diferentes tipos de anomalias. Através da comparação com técnicas clássicas de deteção, como o STL e o LSTM, foi possível aferir a robustez e potencial destas métricas na monitorização da qualidade dos dados.

O segundo estudo introduziu um novo grau de complexidade, aplicando as métricas a um sistema real de bombeamento de água com histórico de falhas operacionais. Para além da aplicação das métricas básicas, este caso destacou-se pela introdução dos índices WQS, LWQS e QSD, que possibilitaram a agregação de múltiplas dimensões da qualidade num único valor interpretável. A utilização destes índices revelou-se eficaz na priorização de sensores críticos e na deteção antecipada de degradações, promovendo uma visão mais holística e funcional da qualidade dos dados em tempo real.

O terceiro estudo representou a consolidação do percurso, com a implementação de um *pipeline* em tempo real, integrando as métricas e funcionalidades testadas previamente. A arquitetura desenvolvida demonstrou ser leve, modular e adaptável, com potencial para aplicação em contextos produtivos reais. A utilização de tecnologias como *Apache Kafka*, *InfluxDB* e *Grafana* possibilitou a ingestão contínua de dados, o cálculo automatizado das métricas de qualidade e a disponibilização de resultados através de *dashboards* interativos. Esta solução revelou-se particularmente eficaz na deteção precoce de falhas, no diagnóstico de comportamentos anómalos e na promoção de uma gestão mais proativa da informação.

Os resultados obtidos reforçaram a importância de integrar mecanismos de avaliação da qualidade dos dados nos sistemas industriais digitais, não apenas como ferramentas de diagnóstico, mas também como alicerces fundamentais para a tomada de decisão baseada em dados. O *pipeline* proposto apresentou-se como uma solução escalável e prática, facilmente integrável em diferentes contextos industriais, promovendo maior confiança, fiabilidade e eficiência nos processos, e contribuindo para a maturidade digital das organizações.

Não obstante os contributos alcançados, foram identificadas algumas limitações relevantes. A definição de regras de consistência baseadas apenas em correlações extraídas dos dados exige validação adicional por parte de especialistas do domínio, sob risco de interpretações incorretas. A dependência da infraestrutura de recolha, nomeadamente em termos de granularidade e estabilidade das fontes de dados, também pode condicionar a aplicabilidade da solução em ambientes industriais menos digitalizados. Por fim, embora a arquitetura tenha sido testada com sucesso em ambientes controlados e com um número limitado de sensores, não foi ainda possível validar o seu desempenho em cenários de larga escala, com múltiplos fluxos paralelos e volumes de dados massivos.

O trabalho desenvolvido nesta dissertação abre espaço a diversas oportunidades de evolução e aprofundamento. Em primeiro lugar, recomenda-se a aplicação da arquitetura a sistemas industriais com características distintas, nomeadamente setores com elevada variabilidade operacional, ciclos produtivos curtos ou maior sensibilidade à latência. A realização de testes com um número superior de sensores e sob diferentes regimes de carga permitirá validar a escalabilidade e a robustez da solução em ambientes de produção reais. Adicionalmente, a integração de especialistas de domínio na definição de regras semânticas e de consistência poderá contribuir para uma maior fiabilidade na deteção de anomalias e para uma contextualização mais precisa dos alertas. Por fim, do ponto de vista técnico, será igualmente pertinente explorar a integração de modelos de aprendizagem para a adaptação dinâmica de parâmetros e pesos nos índices compostos, promovendo um ajustamento mais inteligente às condições operacionais.

Em suma, esta dissertação contribui de forma significativa para o avanço do conhecimento na área da qualidade dos dados no contexto da I4.0, ao propor uma abordagem integrada que alia rigor metodológico, aplicabilidade prática e flexibilidade tecnológica. As soluções desenvolvidas, desde as métricas e índices compostos até à arquitetura completa de monitorização, respondem a um desafio central da transformação digital: garantir que os dados, enquanto ativos críticos, são fiáveis, completos, consistentes e disponíveis em tempo útil.

O *pipeline* concebido e validado representa um passo concreto na direção de sistemas industriais mais inteligentes, resilientes e orientados a dados. Ao incorporar a avaliação contínua da qualidade como parte integrante do ciclo de vida da informação, a solução proposta não apenas deteta problemas de forma proativa, como também fornece indicadores valiosos para a otimização de processos, a redução de custos operacionais e o suporte à manutenção preditiva. A sua arquitetura modular e leve torna-a especialmente adequada a realidades industriais com diferentes níveis de maturidade digital, facilitando a sua adoção progressiva em contextos reais.

Deste modo, considera-se que a dissertação dá uma resposta concreta e sustentada à questão de investigação que a orientou:

De que forma se pode analisar, avaliar e monitorizar, em tempo real, a quali-

dade dos dados em ambientes industriais?

Através do desenvolvimento e validação de soluções que integram a monitorização contínua com a flexibilidade exigida por ambientes industriais heterogéneos, foi possível demonstrar que é viável construir um *pipeline* capaz de garantir visibilidade, fiabilidade e capacidade de reação, permitindo analisar, avaliar e monitorizar a qualidade dos dados em tempo real.

Para além de reforçar a confiança nos dados utilizados na tomada de decisão, este trabalho oferece uma base sólida para investigações futuras, com potencial para adaptação a setores variados e integração com outras camadas analíticas, como modelos de IA ou sistemas de apoio à decisão. O *pipeline* desenvolvido não é apenas uma prova de conceito, mas uma ferramenta concreta que pode ser expandida, refinada e aplicada em larga escala, contribuindo de forma sustentada para o avanço da transformação digital industrial e para a construção de ecossistemas produtivos mais eficientes, autónomos e orientados à qualidade da informação.

Capítulo 6

Resultados Científicos

Este capítulo reúne os artigos científicos desenvolvidos no âmbito desta dissertação. Cada publicação é brevemente descrita e incluí o *link* para o artigo *on-line*, de forma a permitir o seu acesso. Estes resultados constituem um contributo para o avanço do conhecimento na área da qualidade dos dados em ambientes industriais, particularmente no contexto da I4.0.

As publicações cobrem diferentes vertentes do problema, desde a revisão do estado da arte até à aplicação prática de metodologias de avaliação e monitorização da qualidade dos dados em cenários reais. A abordagem adotada permitiu não só aprofundar o enquadramento teórico do tema, como também demonstrar a aplicabilidade e eficácia das soluções propostas em ambientes industriais com diferentes níveis de complexidade. Assim, os artigos aqui apresentados serviram de base para a estruturação da presente dissertação e para a validação dos casos de estudo desenvolvidos, refletindo a evolução e maturação da investigação realizada ao longo do projeto.

Artigo 1 – *Extensible Data Ingestion System for Industry 4.0*

Conferência: *EPIA 2024 Conference on Artificial Intelligence*
Data: Novembro 2024
Link: `https://doi.org/10.1007/978-3-031-73503-5_9`
Autores: Bruno Oliveira, Óscar Oliveira, Teresa Peixoto, Fillipe Ribeiro e Carla Pereira (8)

Resumo: Este artigo propõe uma arquitetura extensível para ingestão e monitorização de dados em ambientes industriais inteligentes, focando-se na recolha, organização e avaliação da qualidade dos dados provenientes de múltiplas fontes. A solução utiliza o *Apache Kafka* como mecanismo central de comunicação entre fontes de dados e serviços de processamento, garantindo fiabilidade e escalabilidade. O sistema é complementado por serviços auxiliares que asseguram o armazenamento em diferentes bases de dados e a visualização em tempo real com Grafana. A arquitetura apresentada destaca-se pela sua flexibilidade, extensibilidade e capacidade de adaptação a diferentes contextos industriais.

Para além de contribuir diretamente para o Capítulo Estado da Arte na área da ingestão

de dados em contextos industriais, esta arquitetura serviu também de base para a implementação do *pipeline* de ingestão e monitorização utilizado no Caso de Estudo 3 (Artigo 5), validado num cenário real de produção.

Artigo 2 – *Real-Time Manufacturing Data Quality: Leveraging Data Profiling and Quality Metrics*

Conferência: *IoTBDS 2025 – International Conference on Internet of Things, Big Data and Security*
Data: Janeiro 2025
Link: `https://doi.org/10.5220/0013242900003944`
Autores: Teresa Peixoto, Bruno Oliveira, Óscar Oliveira e Fillipe Ribeiro (3)

Resumo: Este artigo apresenta uma abordagem para a monitorização da qualidade dos dados em tempo real em ambientes de manufatura inteligente. A solução combina técnicas de *data profiling* com o cálculo de métricas adaptadas às principais dimensões de qualidade, acurácia, completude, consistência e atualidade, permitindo a avaliação contínua de dados provenientes de sensores industriais. O sistema recorre a blocos de tempo de cinco minutos para cálculo incremental das métricas, utilizando percentis dinâmicos e regras derivadas do comportamento histórico dos dados.

O estudo é validado através de um caso simulado no processo de extrusão de plástico, demonstrando a eficácia do método na deteção de anomalias, identificação de falhas em sensores e apoio à tomada de decisão. A integração de tarefas de *data profiling* permite automatizar a deteção de problemas de qualidade, reduzir a intervenção manual e reforçar a fiabilidade dos dados utilizados em sistemas de manutenção preditiva e controlo de processos industriais.

Este artigo serve de base ao desenvolvimento do Caso de Estudo 1 apresentado nesta dissertação.

Artigo 3 – *Data Quality Assessment in Smart Manufacturing: A Review*

Revista: *Systems*, Volume 13
Data: Março 2025
Link: `https://doi.org/10.3390/systems13040243`
Autores: Teresa Peixoto, Bruno Oliveira, Óscar Oliveira e Fillipe Ribeiro (1)

Resumo: Este artigo apresenta uma revisão sistemática sobre a avaliação da qualidade dos dados em ambientes de manufatura inteligente, com foco nos desafios e requisitos impostos pela I4.0. São analisadas diferentes propostas de classificação das dimensões da qualidade dos dados, destacando-se as dimensões de acurácia, completude, atualidade e consistência como as mais relevantes para contextos industriais.

A segunda parte do artigo aprofunda a descrição de métricas para cada uma dessas dimensões, avaliadas com base em critérios como clareza, relevância e controlabilidade. O estudo enfatiza a importância da monitorização contínua da qualidade para garantir a fiabilidade dos dados em tempo real. Também são discutidos os impactos dos dados ocultos

e sugeridas abordagens como *sketching* para otimizar o processamento e reduzir a perda de valor dos dados.

Este artigo serviu como base para a construção do Capítulo Estado da Arte da presente dissertação, oferecendo o enquadramento teórico das dimensões e métricas de qualidade dos dados, bem como critérios de avaliação para cada métrica.

Artigo 4 – *Comparative Analysis of Anomaly Detection Techniques for IoT Time Series Data*

Conferência: *Business and Technology* 2025
Data de Apresentação: Abril 2025
Autores: Teresa Peixoto, Bruno Oliveira, Óscar Oliveira e Fillipe Ribeiro
Info: Com possibilidade de publicação.

Resumo: Este artigo analisa comparativamente diferentes técnicas de deteção de anomalias aplicadas a séries temporais em ambientes IdC, com foco na monitorização em tempo real e na manutenção preditiva. O objetivo é comparar diferentes tipos de deteção de anomalias conciliando rapidez e precisão em sistemas industriais.

O estudo aplica quatro técnicas distintas, uma métrica de acurácia, o algoritmo *t-digest*, a decomposição STL e um modelo LSTM, sobre dados recolhidos por sensores de temperatura durante o processo de extrusão de plástico. Cada técnica é avaliada com base no número de anomalias detetadas, tempo de execução e adequação às duas zonas da arquitetura proposta: a *Rapid Screening Zone*, onde se privilegia a velocidade, e a *In-Depth Analysis Zone*, onde se procura maior sensibilidade.

Os resultados mostram que métodos como a métrica de acurácia e o *t-digest* são altamente eficientes na deteção imediata de desvios extremos, enquanto o modelo LSTM revela elevada capacidade de deteção em profundidade, embora com maior custo computacional. A técnica *STL* surge como um compromisso entre interpretabilidade e robustez.

Este artigo contribui para o Capítulo Estado da Arte e sustenta e expande a discussão do Caso de Estudo 1.

Artigo 5 – *Data Quality Assessment: A Practical Application*

Conferência: *ICIE 2025 – International Conference on Innovation in Engineering*
Data: Junho 2025
Link: `https://doi.org/10.1007/978-3-031-94484-0_42`
Autores: Eliana Costa e Silva, Teresa Peixoto, Óscar Oliveira e Bruno Oliveira (4)

Resumo: Este artigo apresenta a aplicação prática de métricas de qualidade dos dados num sistema real de bombagem de água, utilizando um conjunto de dados com leituras de 52 sensores ao longo de abril de 2018. São avaliadas duas dimensões principais, acurácia e completude, através de métricas específicas aplicadas em blocos de cinco minutos..

O estudo introduz três índices compostos para avaliação global: o WQS, o LWQS e o QSD, permitindo uma análise integrada da evolução da qualidade ao longo do tempo.

Estes índices permitem priorizar sensores e dimensões de acordo com a criticidade e os objetivos do sistema, oferecendo flexibilidade e adaptabilidade à realidade operacional.

Através da análise de episódios reais de falha e recuperação do sistema, os resultados demonstram que os indices propostos conseguem antecipar quebras na qualidade dos dados, reforçando o seu valor como ferramenta de apoio à decisão. Este artigo contribui diretamente para o desenvolvimento do Caso de Estudo 2 desta dissertação.

Artigo 6 – *A Data Quality Pipeline for Industrial Environments: Architecture and Implementation*

Revista: *Computers*, Volume 14
Data: Junho 2025
Link: `https://doi.org/10.3390/computers14070241`
Autores: Teresa Peixoto, Óscar Oliveira, Eliana Costa e Silva, Bruno Oliveira e Fillipe Ribeiro (5)

Resumo: Este artigo propõe uma abordagem modular e extensível para a monitorização da qualidade dos dados em sistemas industriais reais, com foco em ambientes com restrições de latência e elevada frequência de aquisição de dados. A arquitetura desenvolvida integra serviços de ingestão, avaliação da qualidade e visualização, suportados por tecnologias como *Apache Kafka*, *InfluxDB* e *Grafana*. A solução avalia continuamente quatro dimensões de qualidade, acurácia, completude, consistência e atualidade.

O estudo é validado num ambiente de produção real com sensores de temperatura em sistemas de transporte motorizado, demonstrando a capacidade do sistema em identificar anomalias, perdas de dados e desvios em tempo real. Além disso, são avaliados índices compostos, WQS, LWQS e QSD, que permitem acompanhar a evolução da qualidade e priorizar ações com base em sensores e dimensões críticas. A interface visual desenvolvida em *Grafana* facilita a análise contínua e a tomada de decisão operacional.

Este artigo representa a validação final da arquitetura desenvolvida ao longo da dissertação, consolidando os contributos teóricos e práticos dos casos de estudo anteriores. Contribui diretamente para o Caso de Estudo 3.

Bibliografia

[1] T. Peixoto, B. Oliveira, Ó. Oliveira, and F. Ribeiro, "Data quality assessment in smart manufacturing: A review," *Systems*, vol. 13, no. 4, p. 243, 2025.

[2] M. P. Groover, *Fundamentals of modern manufacturing: materials, processes, and systems*. John Wiley & Sons, 2010.

[3] T. Peixoto, B. Oliveira, Óscar Oliveira, and F. Ribeiro, "Real-time manufacturing data quality: Leveraging data profiling and quality metrics," in *Proceedings of the 10th International Conference on Internet of Things, Big Data and Security - IoTBDS*, pp. 56–68, INSTICC, SciTePress, 2025.

[4] E. C. e Silva, T. Peixoto, Ó. Oliveira, and B. Oliveira, "Data quality assessment: A practical application," in *Innovations in Industrial Engineering IV* (J. Machado, J. Trojanowska, K. Antosz, C. P. Leão, L. Knapcikova, and A. Sover, eds.), (Cham), pp. 512–523, Springer Nature Switzerland, 2025.

[5] T. Peixoto, Oliveira, E. Costa e Silva, B. Oliveira, and F. Ribeiro, "A data quality pipeline for industrial environments: Architecture and implementation," *Computers*, vol. 14, no. 7, 2025.

[6] G. Miragliotta, A. Sianesi, E. Convertini, and R. Distante, "Data driven management in industry 4.0: a method to measure data productivity," *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 19–24, 2018.

[7] S. Munir, S. I. Jami, and S. Wasi, "Knowledge graph based semantic modeling for profiling in industry 4.0," *International Journal on Information Technologies & Security*, vol. 12, no. 1, pp. 37–50, 2020.

[8] B. Oliveira, Ó. Oliveira, T. Peixoto, F. Ribeiro, and C. Pereira, "Extensible data ingestion system for industry 4.0," in *EPIA Conference on Artificial Intelligence*, pp. 105–114, Springer, 2024.

[9] A. Goknil, P. Nguyen, S. Sen, D. Politaki, H. Niavis, K. J. Pedersen, A. Suyuthi, A. Anand, and A. Ziegenbein, "A systematic review of data quality in cps and iot for industry 4.0," *ACM Computing Surveys*, vol. 55, no. 14s, pp. 1–38, 2023.

[10] C. Hu, Z. Sun, C. Li, Y. Zhang, and C. Xing, "Survey of time series data generation in iot," *Sensors*, vol. 23, 2023.

[11] S. Tverdal, A. Goknil, P. Nguyen, E. J. Husom, S. Sen, J. Ruh, and F. Flamigni, "Edge-based data profiling and repair as a service for iot," pp. 17–24, Association for Computing Machinery, 2024.

[12] A. Corallo, A. M. Crespino, V. D. Vecchio, M. Lazoi, and M. Marra, "Understanding and defining dark data for the manufacturing industry," *IEEE Transactions on Engineering Management*, vol. 70, no. 2, 2023. pp. 700–712.

[13] D. Kuemper, T. Iggena, R. Toenjes, and E. Pulvermueller, "Valid. iot: A framework for sensor data quality analysis and interpolation," in *Proceedings of the 9th ACM Multimedia Systems Conference*, pp. 294–303, 2018.

[14] R. Krishnamurthi, A. Kumar, D. Gopinathan, A. Nayyar, and B. Qureshi, "An

overview of iot sensor data processing, fusion, and analysis techniques," *Sensors*, vol. 20, no. 21, 2020.

[15] D. Loshin, *The practitioner's guide to data quality improvement*. Elsevier, 2010.

[16] C. Liu, G. Peng, Y. Kong, S. Li, and S. Chen, "Data quality affecting big data analytics in smart factories: Research themes, issues and methods," *Symmetry*, vol. 13, no. 8, 2021.

[17] H. Cheng, D. Feng, X. Shi, and C. Chen, "Data quality analysis and cleaning strategy for wireless sensor networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, 03 2018.

[18] Z. Abedjan, L. Golab, F. Naumann, and T. Papenbrock, "Data profiling," *Synthesis Lectures on Data Management*, vol. 10, pp. 1–154, 11 2018.

[19] S. Rangineni, A. Bhanushali, M. Suryadevara, S. Venkata, and K. Peddireddy, "A review on enhancing data quality for optimal data analytics performance," *International Journal of Computer Sciences and Engineering*, vol. 11, 2023. pp. 51–58.

[20] L. Zhang, D. Jeong, and S. Lee, "Data quality management in the internet of things," *Sensors*, vol. 21, no. 17, 2021.

[21] M. Qasim Jebur Al-Zaidawi and M. Çevik, "Advanced deep learning models for improved iot network monitoring using hybrid optimization and mcdm techniques," *Symmetry*, vol. 17, no. 3, 2025.

[22] H. Y. Teh, A. W. Kempa-Liehr, and K. I.-K. Wang, "Sensor data quality: A systematic review," *Journal of Big Data*, vol. 7, no. 1, 2020.

[23] R. Mahanti, *Data Quality: Dimensions, Measurement, Strategy, Management, and Governance*. ASQ Quality Press, USA, 2019. `https://asq.org/quality-press/display-item?item=H1552`.

[24] C. Batini, M. Scannapieco, *et al.*, *Data and information quality*, vol. 63. Springer, 2016.

[25] "Software engineering Software product Quality Requirements and Evaluation (SQuaRE) Data quality model," standard, International Organization for Standardization, Geneva, CH, Dec. 2008.

[26] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, no. 4, 1996. pp. 5–33.

[27] C. Cichy and S. Rass, "An overview of data quality frameworks," *IEEE Access*, vol. 7, 2019. pp. 24634–24648.

[28] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE communications surveys & tutorials*, vol. 12, no. 2, pp. 159–170, 2010.

[29] C. Ryan, A. Parnell, and C. Mahoney, "Real-time anomaly detection for advanced manufacturing: Improving on twitter's state of the art," *arXiv preprint arXiv:1911.05376*, 11 2019.

[30] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[31] S. Ahmad and S. Purdy, "Real-time anomaly detection for streaming analytics," 2016.

[32] A. Chatterjee and B. S. Ahmed, "Iot anomaly detection methods and applications: A survey," *Internet of Things*, vol. 19, p. 100568, 2022.

[33] B. Sheng, Q. Li, W. Mao, and W. Jin, "Outlier detection in sensor networks," in *Proceedings of the 8th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc '07, (New York, NY, USA), p. 219–228, Association for Computing Machinery, 2007.

[34] L. Lankewicz and M. Benard, "Real-time anomaly detection using a nonparametric pattern recognition approach," in *Proceedings Seventh Annual Computer Security Applications Conference*, pp. 80–81, IEEE Computer Society, 1991.

[35] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250–2267, 2014.

[36] E. Krzysztoń, I. Rojek, and D. Mikołajewski, "A comparative analysis of anomaly detection methods in iot networks: An experimental study," *Applied Sciences*, vol. 14, no. 24, p. 11545, 2024.

[37] T. Dunning, "The t-digest: Efficient estimates of distributions," *Software Impacts*, vol. 7, p. 100049, 2021.

[38] R. Kirkby, "Computing quantiles of functions of the agent distribution using t-digests," *Computational Economics*, vol. 64, no. 2, pp. 1199–1218, 2024.

[39] T. Dunning and O. Ertl, "Computing extremely accurate quantiles using t-digests," *arXiv preprint arXiv:1902.04023*, 2019.

[40] M. Abu-Alhaija and N. M. Turab, "Automated learning of ecg streaming data through machine learning internet of things.," *Intelligent Automation & Soft Computing*, vol. 32, no. 1, 2022.

[41] R. B. Cleveland, W. Cleveland, J. McRae, and I. Terpenning, "Stl: A seasonal-trend decomposition procedure based on loess. 1990," *DOI: citeulike-article-id*, vol. 1435502, 2022.

[42] Q. Wen, J. Gao, X. Song, L. Sun, H. Xu, and S. Zhu, "Robuststl: A robust seasonal-trend decomposition algorithm for long time series," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 5409–5416, 2019.

[43] D. Chen, J. Zhang, and S. Jiang, "Forecasting the short-term metro ridership with seasonal and trend decomposition using loess and lstm neural networks," *Ieee Access*, vol. 8, pp. 91181–91187, 2020.

[44] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The performance of lstm and bilstm in forecasting time series," in *2019 IEEE International conference on big data (Big Data)*, pp. 3285–3292, IEEE, 2019.

[45] A. Verner and S. Mukherjee, "An lstm-based method for detection and classification of sensor anomalies," in *Proceedings of the 2020 5th International Conference on Machine Learning Technologies*, pp. 39–45, 2020.

[46] Y. Kim and K. Lee, "A quality measurement method of context information in ubiquitous environments," in *2006 International conference on hybrid information technology*, vol. 2, pp. 576–581, IEEE, 2006.

[47] A. Karkouch, H. Mousannif, H. Al Moatassime, and T. Noel, "Data quality in internet of things: A state-of-the-art survey," *Journal of Network and Computer Applications*, vol. 73, pp. 57–81, 2016.

[48] T. Cerquitelli, N. Nikolakis, P. Bethaz, S. Panicucci, F. Ventura, E. Macii, S. Andolina, A. Marguglio, K. Alexopoulos, P. Petrali, *et al.*, "Enabling predictive analytics for smart manufacturing through an iiot platform," *IFAC-PapersOnLine*, vol. 53, no. 3, pp. 179–184, 2020.

[49] E. Seghezzi, M. Locatelli, L. Pellegrini, G. Pattini, G. M. Di Giuda, L. C. Tagliabue, and G. Boella, "Towards an occupancy-oriented digital twin for facility management: Test campaign and sensors assessment," *Applied Sciences*, vol. 11, no. 7, p. 3108, 2021.

[50] D. C. Corrales, J. C. Corrales, and A. Ledezma, "How to address the data quality issues in regression models: A guided process for data cleaning," *Symmetry*, vol. 10, no. 4, p. 99, 2018.

[51] J. Byabazaire, G. O'Hare, and D. Delaney, "Using trust as a measure to derive data quality in data shared iot deployments," in *2020 29th International Conference on Computer Communications and Networks (ICCCN)*, pp. 1–9, IEEE, 2020.

[52] S. Sicari, C. Cappiello, F. D. Pellegrini, D. Miorandi, and A. Coen-Porisini, "A security-and quality-aware system architecture for internet of things," *Information Systems Frontiers*, vol. 18, 8 2016. pp. 665-677.

[53] D. Ballou, R. Wang, H. Pazer, and G. Tayi, "Modeling information manufacturing systems to determine information product quality," *Management Science*, vol. 44, 04 1998. pp. 462–484.

[54] J. Lee, B. Bagheri, and H.-A. Kao, "A cyber-physical systems architecture for industry 4.0-based manufacturing systems," *Manufacturing Letters*, vol. 3, pp. 18–23, 1 2015.

[55] E. Oztemel and S. Gursev, "Literature review of industry 4.0 and related technologies," *Journal of intelligent manufacturing*, vol. 31, no. 1, pp. 127–182, 2020.

[56] J. Hover, "Data profiling: What, why and how?," 2016.

[57] O. Azeroual, G. Saake, and E. Schallehn, "Analyzing data quality issues in research information systems via data profiling," *International Journal of Information Management*, vol. 41, pp. 50–56, 2018.

[58] Z. Abedjan, L. Golab, and F. Naumann, "Profiling relational data: a survey," *The VLDB Journal*, vol. 24, pp. 557–581, 2015.

[59] F. Cremer, B. Sheehan, M. Fortmann, A. N. Kia, M. Mullins, F. Murphy, and S. Materne, "Cyber risk and cybersecurity: a systematic review of data availability," *The Geneva Papers on Risk and Insurance - Issues and Practice*, vol. 47, pp. 698–736, 7 2022.

[60] W. Dai, I. Wardlaw, Y. Cui, K. Mehdi, Y. Li, and J. Long, "Data profiling technology of data governance regarding big data: Review and rethinking," vol. 448, pp. 439–450, Springer Verlag, 2016.

[61] A. Bronselaer, "Data quality management: An overview of methods and challenges," pp. 127–141, Springer International Publishing, 2021.

[62] R. Kimball, "Kimball design tip 59 : Surprising value of data profiling," pp. 1–2, 2004.

[63] S. Cirillo, "Data stream profiling: Evolutionary and incremental algorithms for dependency discovery," 2022.

[64] C. Ji, Q. Shao, J. Sun, S. Liu, L. Pan, L. Wu, and C. Yang, "Device data ingestion for industrial big data platforms with a case study," *Sensors*, vol. 16, p. 279, 2 2016.

[65] N. Sawant and H. Shah, *Big Data Ingestion and Streaming Patterns*, pp. 29–42. Apress, 2013.

[66] L. Qiao, Y. Li, S. Takiar, Z. Liu, N. Veeramreddy, M. Tu, Y. Dai, I. Buenrostro, K. Surlaker, S. Das, and C. Botev, "Gobblin," *Proceedings of the VLDB Endowment*, vol. 8, pp. 1764–1769, 8 2015.

[67] M. Irfan and J. P. George, *A Systematic Review of Challenges, Tools, and Myths of Big Data Ingestion*, pp. 481–494. 2022.

[68] P. Vyas, A. Shinde, D. Diwase, and A. Kathole, "Advancements in data ingestion and processing using hadoop," *SSRN Electronic Journal*, 2023.

[69] Óscar Oliveira and B. Oliveira, "An extensible framework for data reliability assessment," pp. 77–84, SCITEPRESS - Science and Technology Publications, 2022.

[70] J. Khan, R. Dalu, and S. Gadekar, "Defects in extrusion process and their impact on product quality," *International journal of mechanical engineering and robotics research*, vol. 3, no. 3, p. 187, 2014.

[71] M. Maity and P. Saha, "Normal distribution," *International Journal of Science and Research (IJSR)*, vol. 12, pp. 298–299, December 2023.

[72] M. Sterjev, "T-digest based alerting systems." LinkedIn Pulse, August 2018. https://www.linkedin.com/pulse/t-digest-based-alerting-systems-marjan-sterjev/, Accessed: March 21, 2025.

[73] C. T. Ho, "Anomaly detection in time series," January 2025. https://blog.jetbrains.com/pycharm/2025/01/anomaly-detection-in-time-series, Accessed: March 20, 2025.

[74] T. Peixoto, "Iot anomaly detection results," 2025. https://doi.org/10.5281/zenodo.15112838, GitHub repository.

[75] A. K. Roy, K. K. Jena, and D. Mohapatra, "Water pump health prediction in industrial wireless sensor network using supervised learning," in *2023 OITS International Conference on Information Technology (OCIT)*, pp. 178–183, 2023.

[76] E. Costa e Silva, Ó. Oliveira, and B. Oliveira, "Enhancing real-time analytics: Streaming data quality metrics for continuous monitoring," in *Proceedings of the 2024 7th International Conference on Mathematics and Statistics*, pp. 97–101, 2024.

[77] M. Prabhudesai, "Conveyor belt system with 3 degrees of freedom," *International Journal for Research in Applied Science and Engineering Technology*, vol. 6, pp. 407–411, 06 2018.

[78] S.-A.-F. Team, "Outlier detection in time series." `https://s-ai-f.github.io/Time-Series/outlier-detection-in-time-series.html`, 2023. Accessed: 2025-05-22.

[79] S. Bhowmik, B. Jelfs, S. P. Arjunan, and D. K. Kumar, "Outlier removal in facial surface electromyography through hampel filtering technique," in *2017 IEEE Life Sciences Conference (LSC)*, pp. 258–261, 2017.