Automating Indicator Validation for Water Utility Benchmarking: A Data-Driven Approach

MSc Thesis Project with the Delft University of Technology and the World Bank Group

Pravesha S.P. Ramsundersingh

Electrical Engineering, Mathematics, and Computer Science
Department of Intelligent Systems within Research Group for Multimedia Computing
Master in Computer Science

Delft, The Netherlands, August 2025

Automating Indicator Validation for Water Utility Benchmarking: A Data-Driven Approach

MSc Thesis Project with the Delft University of Technology and the World Bank Group

Pravesha S.P. Ramsundersingh

to obtain the degree of Master of Science at the Delft University of Technology, to be defended publicly on Friday, August 29, 2025 at 09:00.

Thesis Committee: Dr. Ir. Cynthia C. S. Liem
Associate Professor in the Multimedia Computing Group, Delft University of Technology
Dr. Ir. Mr. Vandana N. S. R. Dwarka
Assistant Professor in the Numerical Analysis Group, Delft University of Technology
Dr. Ir. Tom Viering
Assistant Professor in the Pattern Recognition and Bio-Informatics Group, Delft University of Technology

External Supervisors: Marco Antonio Aguero
Senior Water Supply and Sanitation Specialists, World Bank Group
Dr. Monika Weber-Fahr
Senior Partnerships Advisor, World Bank Group

TU Delft

Automating Indicator Validation for Water Utility Benchmarking: A Data-Driven Approach

Acknowledgements

This thesis has been both a craft and a compass, refining my abilities as a researcher while sharpening the values I hope will define my path ahead: innovation, integrity, and impact. Throughout this process, I have been guided and inspired by an extraordinary network of people, each leaving their own mark on the work you see here.

I am deeply grateful to my thesis committee, **Dr. Cynthia C. S. Liem, Dr. Vandana N.S. R. Dwarka, and Dr. Tom Viering**, whose expertise have shaped not only the course of this thesis, but also my growth as a scholar.

First and foremost, I am sincerely thankful to my supervisor, **Dr. Cynthia C. S. Liem**, whose accomplishments stand at the very intersection of computer science, diversity, and ethics – and who has long been an inspiration not only to me, but to many others and the generations of researchers to come. Your belief in the power of research to create societal change has influenced the norms and values I hope to carry forward in my own career, and it has been an honour to learn under your guidance.

To my daily supervisor, **Dr. Vandana N.S. R. Dwarka** – taking on this supervision in your own time was an act of trust that I cannot fully put into words. Your faith in me, and your willingness to invest so much of yourself into this project, will always stay with me.

To **Dr. Tom Viering**, who has seen me close both my Bachelor's and Master's chapters, thank you for your insight and for being part of these milestone moments in my academic life.

I wish to extend my heartfelt thanks to the World Bank's New International Benchmarking Network for Water and Sanitation Utilities (NewIBNET) team. To **Dr. Monika Weber-Fahr**, for opening the door to this collaboration and guiding me through the process; **Marco Antonio Aguero**, for your mentorship and for giving me the space to grow; **Guilherme Almeida Monteiro**, for your exceptional expertise and patience in answering every question I could possibly ask; and to **Ana Badhofer** and **Berta Macheve**, for your contributions in joint meetings and for moulding NewIBNET into what it is today.

For their academic guidance and specialised insights, I thank **Prof. Dr. Klaas Schwartz**, whose advice strengthened the foundations of this work; **Dr. James Hutton**, for enriching the ethical perspective of my thesis; and **Marijn Roelvink**, for helping me understand how data challenges truly unfold within sectoral realities. I also extend my gratitude to **Mr. Agus Sunara**, Advisor to the Board of Directors of PERPAMSI, for providing additional data that allowed me to refine my analysis and add a valuable new

dimension to this thesis.

My sincere thanks go also to **Dr. G.F. (Tina) Nane**, whose generosity and example remind me every day of the importance of paying it forward; and to **Dr. Jorge Martinez** and **Sanne Alblas**, for years of steadfast support across my personal, academic, and professional life.

Beyond academia, I have been blessed with friendships and communities that have been my anchor.

Aleksander Buszydlik, thank you for the years of collaboration, from our Faculty Student Council days to your guidance as a friend and as a thesis role model.

To the Librae Network – **Samantha L. Gabree, Saraf Nawar, Lonneke J. Bierens, Bente F. Wildeboer, and Anna Belenguer Martí** – thank you for filling this year with laughter, bold ideas, and the joy of building something meaningful together. You embody what it means to be unapologetically ambitious, and it has been a privilege to create and grow alongside you.

To my best friends, **Serban A. Alexandru** and **Isabel D.M.T. Zijlmans**, for years of friendship, and for always being my solid rock through every high and low.

Finally, to my family – the truest foundation of everything I have achieved.

To my mother, **Oesha S. I. Thakoerdin** – the embodiment of the person I aspire to be. You carry the weight of many worlds at once, balancing them with a strength that never wavers and a grace that never fades. The wisdom you have passed on has become the foundation of who I am, anchoring me in values that shape every decision I make. Your love has been my safe harbour through every chapter of my life.

To my father, **Atem S. Ramsundersingh** – your wisdom has been like a steady current, carrying new lessons when I least expect them. Your mastery in business and technology, paired with a grounded spiritual strength, has been my guiding force and a source of courage. I carry your clarity, perseverance, and principles into all that I do.

To my brother, **Prashant G.A. Ramsundersingh** – my closest companion and, in many ways, my twin in spirit. You fill my life with laughter that lightens the heaviest days, and with a curiosity that turns even ordinary moments into adventures. You remind me that joy is not a distraction from life's serious work, but an essential part of it.

This thesis is as much yours as it is mine.

Pravesha S.P. Ramsundersingh
Delft, August 2025

Abstract

Water flows through every aspect of life, yet the story of its delivery is only as reliable as the data that records it. In global benchmarking, such data is often uneven, incomplete, and rarely subjected to systematic validation, allowing anomalies to shape perceptions of performance before they are critically examined. This thesis addresses that gap by developing and evaluating a multi-stage, data-driven anomaly detection framework within the World Bank's New International Benchmarking Network for Water and Sanitation Utilities (NewIBNET), situated at the intersection of data science, water governance, and digital ethics.

The framework weaves together four complementary layers – structural validation, rule-based logical checks, peer comparison, and weighted prioritisation – transforming anomaly detection from a surface-level cleaning task into a structured process of active quality assurance. Developed through an iterative, expert-informed process, it is reproducible and adaptable, balancing statistical rigour with the contextual realities of the water sector so that each flag raised carries both analytical credibility and practical relevance.

Applied to the 2022–2024 NewIBNET dataset, the framework is assessed through robustness checks, a national case study of Indonesian utilities, and an expert survey. Results show that it improves anomaly interpretability, limits the propagation of flawed data into comparative analyses, and reduces review time from 75 hours to under 2 minutes – earning unanimous expert endorsement for operational deployment.

By translating the principles of automated, ethically grounded validation into a scalable methodology, this work advances the state of practice in anomaly detection for data-scarce sectors. In shifting from *red flags to real solutions*, it demonstrates how automated validation can turn detection into action, building trust where data meets water, and enabling more transparent, equitable decisions in global water governance.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Water weaves through every aspect of life – from the health of individuals and communities to the stability of economies and the resilience of ecosystems[1] (World Bank, 2023; United Nations, 2025). As pressures from urbanisation, climate change, and demographic shifts intensify, so too does the strain on water systems around the world[2] (World Health Organisation, 2023). Managing these systems effectively demands more than engineering expertise or infrastructure investment; it requires a deep and evolving understanding of how water is delivered, measured, and valued in diverse contexts.

Over the past two decades, this understanding has become central to global development efforts. With initiatives like the Sustainable Development Goals (SDGs)[3], particularly SDG 6 on clean water and sanitation, the international community has committed not only to expanding access, but also to monitoring progress and ensuring accountability. Yet beneath these ambitions lies a quieter, more complex challenge: knowing whether the data we rely on actually reflects reality.

In many parts of the world, data on water services exists in fragments. It is collected in different formats, under varying definitions, and often with limited capacity for verification. These gaps are not merely technical – they shape how decisions are made, how resources are allocated, and which communities receive attention (World Bank, 2011). In this landscape, **data is not neutral**. It carries weight, influences trust, and increasingly serves as a proxy for institutional performance (Tsai, 2025).

Emerging technologies offer new tools for addressing this chal-

[1] **United Nations (UN):** *"Water is essential not only to health, but also to poverty reduction, food security, peace and human rights, ecosystems and education."* (United Nations, 2025)

[2] **World Health Organisation (WHO):** *"In 2021, over 2 billion people live in water-stressed countries, which is expected to be exacerbated in some regions as result of climate change and population growth."*(World Health Organisation, 2023)

[3] **Sustainable Development Goals (SDGs):** A set of 17 global objectives established by the United Nations to promote peace, prosperity, and sustainability by 2030. See more: `https://sdgs.un.org/goals`

lenge. Data science can uncover hidden patterns, detect inconsistencies, and support more efficient validation workflows. But the value of these tools depends on how they are embedded into the systems they aim to improve. If used without transparency or context, automation risks reinforcing the very asymmetries it seeks to resolve. If designed carefully with room for human insight, ethical safeguards, and local nuance, it can strengthen accountability and scale impact.

It is within this tension between responsibility and opportunity that this research is situated. In response to a real-world assignment from the **World Bank**[4], this thesis explores how mathematical models might support the transformation of a legacy benchmarking system – the **New International Benchmarking Network for Water and Sanitation Utilities** (**NewIBNET**)[5] – into a more robust, semi-automated platform. The long-term objective of NewIBNET is to establish itself as a universally accessible and intuitive benchmarking tool that fosters transparency and drives continuous improvement in the global water sector[6]. Achieving this ambition requires not only broader participation but also greater confidence in the quality and comparability of submitted data. As such, the task at hand involves rethinking how data is validated before it enters the system – especially in the absence of strong ground truth. The question at the heart of this thesis is:

> *How can data-driven mathematical models enhance validation and benchmarking of water utility indicators while ensuring reliability, decision-making integrity, and ethical transparency?*

Concretely, this work investigates methods for assessing the quality of submitted performance data before it is incorporated into a central benchmarking repository. With limited ground truth data, ensuring reliability becomes an inherently complex challenge that demands a careful synthesis of technical modelling, contextual awareness, and ethical scrutiny. The NewIBNET platform serves as the principal case study through which these broader questions are explored.

By engaging with this question, the thesis contributes to an ongoing conversation about how digital systems can serve public good – not just efficiently, but responsibly.

[4] **The World Bank:** Read more at https://www.worldbank.org/ext/en/home

[5] **NewIBNET:** Read more at https://newibnet.org/

[6] *"In the past two decades, IBNET has been recognised for its achievements in: Establishing standard performance indicators that reflect utility performance, promoting transparency by publishing the indicators of utilities that participated in the IBNET network, and developing and using benchmarking methodologies that allow for comparing otherwise very different utilities with each other."* Read more: https://newibnet.org/about-us

1.2 Contributions

This thesis emerges from a practical challenge: to explore how legacy validation workflows in global water utility benchmarking can be reimagined through the lens of modern data science, without losing sight of institutional trust, ethical transparency, and the complexity of public infrastructure systems. Rather than treating automation as a purely technical upgrade, this work engages deeply with what it means to validate data in contexts where accuracy, accountability, and equity are all at stake.

The foundation of this research lies in a systematised literature review (Chapter 3) that brings together perspectives from three domains rarely examined in concert – data science, performance benchmarking in water governance, and digital ethics. This synthesis reveals not only the methodological gaps in current validation practices, but also the conceptual tensions that emerge when data science enters high-stakes decision environments. These findings inform the formulation of research questions that aim to be both technically feasible and socially grounded.

From this basis, the thesis constructs a design methodology (Chapter 4) that aligns institutional needs with analytical depth. The current NewIBNET system is also dissected to uncover its structural limitations (Chapter 5). In doing so, the work redefines *data integrity* as not just statistical correctness, but as a property shaped by consistency, traceability, and meaningful feedback loops.

A comparator-based framework (Chapter 6) is investigated, moving beyond universal thresholds to account for heterogeneity through meaningful peer group comparisons. This is coupled with scoring and prioritisation logic (Chapter 7) that addresses a more subtle problem: not just *whether* a data point deviates, but *how much it matters*, and to whom. In designing this system, special attention is paid to the interpretability of outputs – ensuring that the signals it produces can be understood, interrogated, and, when necessary, contested.

A technical evaluation (Chapter 8) assesses the system's ability to enhance data review processes while supporting human interpretability and institutional confidence. This is followed by a broader reflection (Chapter 9) on the ethical, political, and sectoral dimensions of deploying automated decision-making tools in global public infrastructure.

These contributions are intended not just as a response to a specific assignment, but as a step toward a broader research agenda where *data meets water* in ways that are responsive to the complex world they aim to serve.

1.3 Structure

This thesis is organised into ten chapters. Chapter 1 introduces the broader context, motivation, and central research question. Chapter 2 lays the foundation by explaining key concepts related to water governance, benchmarking, and data quality, and provides a detailed overview of the current NewIBNET pipeline. Chapter 3 presents a systematised literature review, synthesising perspectives from data science, infrastructure benchmarking, and digital ethics to frame the problem space. Chapter 4 outlines the design, including the refinement of key terms, research sub-questions, and methodological approach. Chapter 5 examines the current system through the lens of data integrity, highlighting structural and statistical challenges in preparing data for validation. Chapter 6 explores context-aware modelling approaches that use utility metadata to detect deviations in performance across diverse submissions. Chapter 7 introduces a severity-based scoring framework to translate statistical anomalies into prioritised benchmarking insights. Chapter 8 provides a technical evaluation of the system's performance, robustness, and alignment with expert expectations. Chapter 9 then reflects on the broader ethical, political, and sectoral implications of automating anomaly detection in global water utility benchmarking. Finally, Chapter 10 concludes the thesis by unifying key findings and outlining future directions at the intersection of data-driven validation and responsible public-sector automation.

2 Background & System Architecture

This thesis focuses on the exploration of an *automated key performance indicator* (*KPI*) *data validation model* to improve the *efficiency of reviewing benchmarking data* submitted by *water utilities*, using the *World Bank's NewIBNET* system as a case study environment, in a way that supports *reliability, decision-making integrity, and ethical transparency*. To build the foundation for this work, Section 2.1 Background unpacks the key terms and concepts embedded in the research question, providing essential definitions and contextual framing that will support later system analysis. Section 2.2 System Architecture then presents an overview of the current NewIBNET data pipeline.

2.1 Background

This section focuses on the key terms and concepts referenced in the main research question, providing the necessary institutional and operational context for the chapters that follow. Section 2.1.1 defines water utilities, their role in ensuring reliable water supply, and the importance of performance tracking. Section 2.1.2 introduces the World Bank and the evolution of IBNET, highlighting its role in global water utility benchmarking and the transition to the NewIBNET system. Section 2.1.3 discusses the KPI benchmarking process. Finally, Section 2.1.4 examines the need for automated validation, addressing inefficiencies in the existing system and the necessity of scalable, data-driven solutions.

2.1.1 Water Utilities

> A ***water utility*** is a governmental, municipal, or private entity responsible for sourcing, treating, and distributing safe and reliable drinking water to households, businesses, and industries (North Carolina Water Service, 2024; Glickman, 2014).

Water is a fundamental element of life – essential for human survival, economic development, and environmental sustainability (World Bank, 2023). Yet, despite its necessity, millions[7] worldwide still face water scarcity and inadequate sanitation. As population growth and changing agricultural land use intensify pressures on global water resources, the efficient management of water supply systems has become more crucial than ever. Water utilities sit at the heart of this challenge, bridging the gap between natural water sources and the communities that rely on them. Through infrastructure development, technological innovation, and regulatory governance, they ensure a steady supply of potable water while promoting the long-term sustainability of water utilities and the broader water supply system. Their role extends beyond providing a basic service; they contribute to economic growth, public health, and resilience against water-related crises.

However, defining the exact role and function of water utilities is not always straightforward. Research papers and policy frameworks (North Carolina Water Service, 2024; Glickman, 2014) often present varying definitions, reflecting differences in regional needs, governance structures, and regulatory environments. To establish a common understanding and ensure accountability, utilities themselves track ***key performance indicators*** (***KPIs***)[8] – quantitative measures used to assess their own progress over time and identify improvements or areas needing attention.

Yet in practice, the reality behind these indicators is often uneven. While utilities can reliably use KPIs to track their own performance, definitions and measurement practices vary widely across contexts, making international or cross-utility comparisons more difficult. What is measured, how it is measured, and with what degree of accuracy can differ substantially – shaped by contextual constraints, institutional capacity, and evolving interpretations of what *good* performance means. These discrepancies are not necessarily flaws in intent, but reflections of the practical limitations util-

[7] **United Nations (UN):** *"The global urban population facing water scarcity is projected to double from 930 million in 2016 to 1.7–2.4 billion people in 2050."* (United Nations, 2025)

[8] Examples include population service size, water quality metrics, and operational efficiency.

ities face on the ground. While broader alignment across utilities may be an important long-term goal, the more immediate question – and the one this thesis takes up – is how regulators, governments, supporting agencies, and utility leaders themselves might still extract meaningful signals from data that is inevitably imperfect.

Rather than assuming consistency or completeness, this work begins with the data as it arrives: sometimes noisy, sometimes partial, but still carrying traces of operational reality. The challenge is to design a validation approach that recognises uncertainty without defaulting to distrust, and that strengthens institutional benchmarking without demanding perfection. In doing so, this thesis aims to contribute not only a technical solution, but a perspective – one that embraces imperfection as a starting point, not a dead end.

2.1.2 The World Bank & NewIBNET

> The ***World Bank*** is an international financial institution committed to reducing poverty and promoting sustainable development by providing financial and technical assistance to countries worldwide. It supports projects that aim to stimulate economic growth, strengthen infrastructure, and expand access to essential services – including water and sanitation, which remain core to its development agenda.

The World Bank is a lending institution with a development-oriented mission, offering loans and credits – often with interest – to ensure the sustainability of its financial operations. This role underscores the importance of designing data and benchmarking systems that are not only technically robust and context-sensitive, but also institutionally aligned with long-term viability.

Recognising the unique challenges faced by the water and sanitation sector, the World Bank[9] established the ***International Benchmarking Network for Water and Sanitation Utilities (IBNET)***[10] in 1994. Unlike firms in competitive markets, which are continuously driven to improve by market forces, water utilities often operate in monopolistic and highly resource-constrained environments, shielded from direct competition. As a result, while some utilities proactively enhance their performance, others stagnate, falling behind best practices. This disparity has far-reaching consequences – only well-managed and financially stable utilities can effectively expand services, respond to urban growth, and ensure safe wastew-

[9] **World Bank:** Read more: https://www.worldbank.org/ext/en/who-we-are

[10] **IBNET:** Read more: https://www.ib-net.org/about-us/

ater management.

IBNET emerged as a global benchmarking initiative, creating a standardised framework for assessing utility performance. By collecting and comparing data across utilities, the initiative aimed to increase transparency and encourage best practices. It provided key stakeholders – utility managers, regulators, and policymakers – with the necessary insights to track and guide sector improvements. Establishing KPIs became central to this effort, allowing utilities to measure progress and governments to refine policies.

With over 3,000 participating utilities across more than 150 countries, IBNET evolved into one of the world's most extensive water utility benchmarking databases. Yet, as its scale grew, so did the challenges of ensuring data quality. The benchmarking process relied heavily on manual data collection, which was slow, inconsistent, and increasingly difficult to manage at scale. As technology advanced, it became evident that a more efficient and automated system was needed to maintain the reliability of data.

These challenges led to the development of the ***NewIBNET***[11] in 2021, a modernised platform designed to improve data validation and streamline benchmarking. While preserving the benchmarking objectives of its predecessor, NewIBNET introduces real-time consistency checks, customisable reports, and interactive dashboards, allowing for more dynamic data analysis. The platform also promotes a *Community of Practice*[12], enabling utilities to connect with industry peers and share insights to adopt best practices tailored to their operational context.

By transitioning to NewIBNET, the World Bank has strengthened its commitment to data-driven decision-making in the water sector. The enhanced system ensures that benchmarking remains both scalable and effective, reinforcing efforts to improve service delivery, expand access, and build a more resilient global water infrastructure.

2.1.3 Efficiency in Benchmarking Data Process

In its current form, the NewIBNET system follows a structured but highly manual review process – one that must contend with variation, inconsistency, and missing information at scale. These imperfections place a considerable burden on the reviewing workflow and raise the question of how limited institutional resources can be used most effectively.

The NewIBNET data process is structured into three main com-

[11] **NewIBNET:** Read more: https://newibnet.org/about-ibnet

[12] *"Community of Practice lets you connect with other utilities in on a dedicated platform where you can learn from their success and challenges."* Read more: https://newibnet.org/utilities

ponents within its process pipeline[13]: **Data Mobilisation**, **Data Review**, and **Data Visualisation**. **Data Mobilisation** involves the online collection of benchmarking data, where each participating utility submits its KPI and Management Practices[14] data from the previous fiscal year. Once submitted, the process moves to the **Data Review** stage, where a single reviewer within the NewIBNET team examines the raw data for over 250+ utilities, resulting in a total of an estimated 75 hours[15] of manual checking annually. Finally, the validated data advances to the **Data Visualisation** stage, where insights are generated for analysis and benchmarking. A key takeaway from the **Data Review** stage is its highly manual nature, requiring meticulous checks of each entry – a process that becomes increasingly time-consuming as the dataset grows.

The ***efficiency of reviewing benchmarking data***, as stated in the thesis focus here, refers to the ability to streamline the manual validation process by integrating automated checks that identify potential inconsistencies. Rather than relying on exhaustive line-by-line reviews, the system could prioritise entries for human attention based on algorithmic signals, enabling a more targeted, interpretable, and ultimately scalable validation process.

The NewIBNET team plays a vital role in maintaining and improving this benchmarking system. By continuously incorporating feedback from water utilities, reviewers, and other stakeholders, they aim to refine the platform, uphold high standards, and ensure its usability across diverse contexts.

2.1.4 The Need for Automated Validation

A fundamental challenge highlighted in Section 2.1.3 is the reliance on a single reviewer to manually assess and validate hundreds of data entries submitted to NewIBNET each year. As the number of participating water utilities grows, so does the volume of data, creating a scalability issue – not only in the size of the dataset but also in the increasing workload placed on the reviewer. This manual-intensive approach introduces the risk of fatigue, making it more likely that inconsistencies, errors, or missing data may slip through the validation process, ultimately compromising the quality and reliability of benchmarking data.

To address this, the concept of automated validation offers a more sustainable solution. ***Automation***[16] in this context refers to reducing the need for tedious, repetitive human analysis by implementing a code-based approach to flagging potential inconsisten-

[13] **NewIBNET Pipeline:** An in-depth overview is provided in Section 2.2.

[14] **Management Practices:** A separate section of the NewIBNET survey containing multiple-choice questions on utility management practices (non-numeric data).

[15] An expert reviewer at NewIBNET estimated that manual review requires approximately 15–20 minutes per utility.

[16] **Harvard Business Review (HBR):** *"Automation reduces the repetitive and monotonous tasks humans have to do by relegating those tasks to software."* (Perez, 2023)

cies. Rather than requiring a reviewer to manually inspect every raw entry, an automated system could optimise the current review process. ***Validation***[17] refers to the systematic process of verifying benchmarking data for accuracy, completeness, and consistency against predefined rules, ensuring its reliability before it informs decision-making.

Beyond improving efficiency, an automated approach must also ensure that reliability, decision-making integrity, and ethical transparency are preserved. ***Reliability***[18] pertains to maintaining consistency and accuracy throughout the validation process, ensuring that flagged entries are genuinely indicative of anomalies rather than the result of faulty detection mechanisms. ***Decision-making integrity***[19] emphasises the importance of providing reviewers with unbiased data to support well-informed and responsible decisions, including determining which data points are valid and which require revision. ***Ethical transparency*** involves designing a system that is clear, understandable, and fair, where the criteria and processes used to flag data are openly communicated, promoting trust and accountability among all stakeholders (Radanliev, 2024).

This is particularly crucial given that industry benchmarks in water utility performance evaluation remain underexplored. Establishing automated validation mechanisms could not only enhance the accuracy of benchmarking but also contribute to filling a critical research gap[20] in the field.

This thesis lays the groundwork for optimising data validation while reinforcing trust and transparency in global water utility benchmarking. By **investigating a data-driven, automated anomaly flagging approach, this work examines how a validation model can be designed to be both technically robust, and sensitive to the socio-political and ethical dynamics influencing stakeholders** – from reviewers to policymakers and, ultimately, end water consumers.

2.2 System Architecture

Having established the foundational terms and institutional context, this section examines the existing operational structure of the NewIBNET system in greater depth.

[17] **Airbyte:** *"Data validation is a systematic process that verifies data accuracy, completeness, and consistency against predefined rules and business logic before it enters decision-making workflows."* (Kutz, 2025)

[18] **Monte Carlo Data:** *"Data reliability is the degree to which data remains accurate, complete, and consistent over time and across various conditions."* (Moses, 2025)

[19] **International Business Machines (IBM):** *"Data integrity is crucial for organizations to trust the data they use for decision-making, as well as to comply with regulatory requirements."* (Jones, 2023)

[20] Expanded upon further in Chapter 3 Literature Review.

2.2.1 Understanding the Existing NewIBNET Pipeline

A fundamental starting point in designing any automated system is understanding the current process it intends to support or replace. Early meetings with the NewIBNET team focused on mapping the existing architecture and identifying critical system requirements. These initial discussions made clear that a nuanced understanding of the end-to-end data flow was essential for pinpointing opportunities for automation, standardisation, and improvement.

The NewIBNET pipeline operates through three main stages shown in Figure 2.1: **Data Mobilisation**, **Data Review**, and **Data Visualisation**, with a short pre-mobilisation step that initiates the cycle.

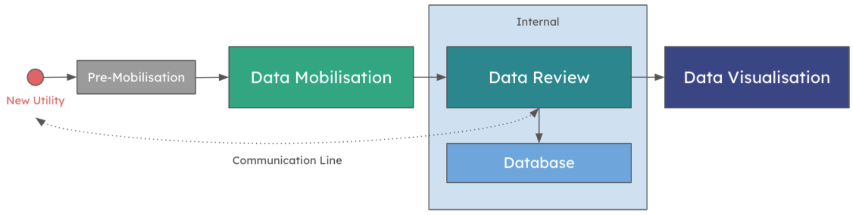


Figure 2.1: *This diagram illustrates the current system architecture of the NewIBNET platform. The process begins with the entry of a registered utility (red dot), which first passes through the pre-mobilisation stage (grey rectangle), followed by data mobilisation (light green rectangle). It then enters the data review stage (dark green rectangle), an internal step visible only to the NewIBNET team, after which the data is entered into the central database (light blue rectangle). A dotted line indicates the communication channel back to the utility. The final stage involves data visualisation (dark blue rectangle).*

This thesis primarily intervenes within the **Data Review** stage, where the core technical design and implementation will take place. However, both the **Data Mobilisation** and **Data Visualisation** stages are examined briefly to understand how upstream data structures influence validation needs, and how downstream benchmarking relies on high-quality inputs. As such, the system is approached holistically, even as the central contribution focuses on optimising the review logic itself.

Pre-Mobilisation

Each utility receives a yearly notification to submit data via the online NewIBNET portal. Access is secured via login credentials, and submissions are one per utility per year. Contact details are collected for follow-up if clarifications are needed, creating accountability and enabling direct validation if unusual patterns are detected.

Data Mobilisation

Utilities complete a structured online survey, with some basic input constraints[21]. While not all data fields are mandatory, the form includes precise definitions[22] to guide respondents and promote consistency. However, deeper validation rules are still limited at this stage.

Data Review

Submitted raw data is manually reviewed by a single expert using a custom-built Excel[23] tool. Management Practice data is not considered within this stage. The reviewer downloads raw submissions and applies rule-based checks using their own formulas and benchmark logic. If unusual results are flagged, the reviewer communicates directly with the utility to request clarification or corrections, ensuring a feedback loop before data is finalised. Approved data is uploaded to secure internal servers at the World Bank, where confidentiality is further protected. Raw data is not made publicly visible. However, there is currently no system in place to log flagged entries or track which data required follow-up.

Data Visualisation

Aggregated results are published on the NewIBNET website dashboard[24].

2.2.2 Survey Setup and Indicator Design

Building on the overview of the NewIBNET system, this section examines the **Data Mobilisation** stage, where the survey[25] instrument serves as the primary interface for annual performance data submission by utilities. This stage shapes the upstream quality of information from which KPIs are derived.

The survey captures a broad spectrum of operational and service dimensions, including basic utility information, service profiles, drinking water coverage, customer interaction, workforce composition, and financial data. While most sections require full completion, those concerning sanitation and wastewater are only relevant to a subset of utilities, introducing heterogeneity in data coverage. Basic data typing rules are enforced, but complex cross-field validations are absent, allowing structural gaps and inconsistencies to arise.

[21] Examples include numeric formats and logical dependencies such as one value not exceeding another.

[22] The default language for all survey questions and expected entries is English.

[23] **Microsoft Excel** is a spreadsheet software widely used for data entry, analysis, and visualisation in benchmarking and reporting. View at: `https://www.microsoft.com/en-us/microsoft-365/excel`

[24] **NewIBNET Dashboard:** View at: `https://newibnet.org/utility-dashboard`

[25] **NewIBNET Survey Sheet:** This definition sheet – detailed in Appendix A.6 – includes the name, question, datatype, unit of measurement, and relevant notes for each field in the survey.

Following data submission, **15 core performance indicators** are calculated from the raw inputs as shown in Figure 2.2.

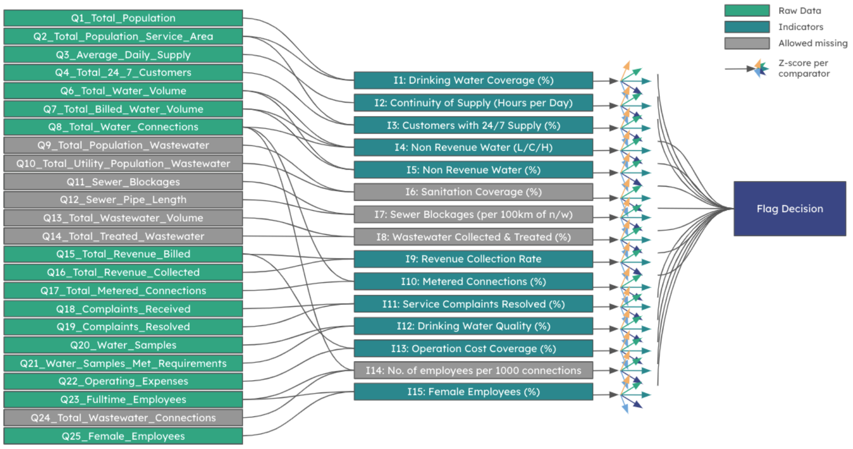

Q1_Total_Population
Q2_Total_Population_Service_Area
Q3_Average_Daily_Supply
Q4_Total_24_7_Customers
Q6_Total_Water_Volume
Q7_Total_Billed_Water_Volume
Q8_Total_Water_Connections
Q9_Total_Population_Wastewater
Q10_Total_Utility_Population_Wastewater
Q11_Sewer_Blockages
Q12_Sewer_Pipe_Length
Q13_Total_Wastewater_Volume
Q14_Total_Treated_Wastewater
Q15_Total_Revenue_Billed
Q16_Total_Revenue_Collected
Q17_Total_Metered_Connections
Q18_Complaints_Received
Q19_Complaints_Resolved
Q20_Water_Samples
Q21_Water_Samples_Met_Requirements
Q22_Operating_Expenses
Q23_Fulltime_Employees
Q24_Total_Wastewater_Connections
Q25_Female_Employees
I1: Drinking Water Coverage (%)
I2: Continuity of Supply (Hours per Day)
I3: Customers with 24/7 Supply (%)
I4: Non Revenue Water (L/C/H)
I5: Non Revenue Water (%)
I6: Sanitation Coverage (%)
I7: Sewer Blockages (per 100km of n/w)
I8: Wastewater Collected & Treated (%)
I9: Revenue Collection Rate
I10: Metered Connections (%)
I11: Service Complaints Resolved (%)
I12: Drinking Water Quality (%)
I13: Operation Cost Coverage (%)
I14: No. of employees per 1000 connections
I15: Female Employees (%)
Flag Decision
Raw Data
Indicators
Allowed missing
Z-score per comparator

Figure 2.2: *This illustrates the transformation pathway from raw numerical inputs (left) – where raw data questions are abbreviated as Q# and grey rectangles indicate optional data entries – through derived indicators (centre), abbreviated as I#, and into the comparator logic that determines final flagging outcomes (right).*

For analytical clarity, these indicators are organised into five semantic categories, reflecting a logical grouping derived from the dataset review:

1. Water Access & Quality Performance
   - I1: Drinking Water Coverage (%)
   - I4: Non Revenue Water (L/C/H)
   - I5: Non Revenue Water (%)
   - I10: Metered Connections (%)
   - I12: Drinking Water Quality (%)
2. Customer Service Performance
   - I2: Continuity of Supply (Hours per Day)
   - I3: Customers with 24/7 Supply (%)
   - I11: Service Complaints Resolved (%)
3. Workforce Metrics
   - I14: No. of employees per 1000 connections
   - I15: Female Employees (%)
4. Sanitation & Wastewater Performance
   - I6: Sanitation Coverage (%)
   - I7: Sewer Blockages (per 100km of n/w)
   - I8: Wastewater Collected & Treated (%)
   - I14: No. of employees per 1000 connections

5. Financial Performance
    - I9: Revenue Collection Rate
    - I13: Operation Cost Coverage (%)

These categories will facilitate a structured examination of how performance data is interpreted and compared in the review process.

2.2.3 Comparison Criteria

Once core indicators are derived from the submitted survey data, they are assessed for plausibility. This decision is guided by a set of ***static comparators*** defined as predefined peer groupings that help reviewers determine whether a utility's performance falls within an acceptable range. These comparators[26] serve as anchors during the **Data Review** stage, aiding in the identification of outliers or inconsistencies.

Currently, three primary static comparator types are used across the review process. These comparators are constructed from historically submitted NewIBNET data and remain unchanged throughout the review period.

1. **Global Average:** A single average value is computed across all utilities for each indicator. This serves as a general reference point, regardless of a utility's features[27].
2. **Population Service Size:** Utilities are grouped based on the size of the population they serve. Indicators are then compared within these predefined population categories, offering a more nuanced comparison.
3. **Number of Water Service Connections:** A similar stratification exists based on the total number of water service connections, aiming to cluster utilities with comparable operational scale.

A second category of comparators, referred to in this thesis as ***dynamic comparators***, is also defined. This is a conceptual distinction for the purposes of this work only, and is not implemented in any form within the current NewIBNET process. Unlike static comparators, dynamic comparators are envisioned as real-time or regularly updated reference groupings, potentially drawing on external datasets beyond the NewIBNET repository. In principle, such comparators could enable continuous recalibration of acceptable ranges, improving contextual relevance and responsiveness, particularly in fast-evolving or data-rich utility environments.

[26] To avoid conceptual ambiguity, this thesis distinguishes between the terms *benchmark*, used to denote public-facing reference values, and *comparator*, used to describe internal comparisons within the dataset.

[27] For example: a utility's size, region, or income level.

2.2.4 System Assessment

While the current NewIBNET pipeline provides a valuable structure from data collection to public dissemination, several key areas within the **Data Review** stage present opportunities for strengthening. At present, the process lacks a codified validation protocol, relies on inconsistent comparator logic, and depends on an undocumented Excel-based tool – factors that pose challenges for scalability, transparency, and long-term institutional resilience. Furthermore, the reliance on global averages as a comparator, while practical, may overlook important contextual differences between utilities, potentially introducing unintended bias into performance assessments.

This thesis seeks to address these limitations by proposing a data-driven validation framework grounded in statistical profiling, comparator-based modelling, and ethically informed prioritisation logic. In doing so, it aims to support the evolution toward more adaptive, transparent, and context-aware benchmarking systems that uphold both technical rigour and institutional trust.

> **Summary:** The current NewIBNET system operates through a structured yet manual pipeline, driven by survey-based data mobilisation, individually applied review logic, and static comparator benchmarks. Though effective in facilitating basic validation, the system lacks the standardisation, scalability, and automation needed for consistent cross-utility assessment. These structural limitations lay the groundwork for the improvements proposed in subsequent chapters.

3

Literature Review & Thematic Foundations

This chapter presents a structured literature review aimed at establishing a comprehensive foundation for developing an automated data validation model. The review is organised into four sections: Setup and Protocol in Section 3.1, which outlines the methodology used to gather and categorise relevant literature; Thematic Content Analysis in Section 3.2, which synthesises findings across the Technical, Water, and Ethical dimensions; Key Insights from the Literature Review in Section 3.3, which contextualises the findings and identifies research gaps; and Limitations of the Literature Review in Section 3.4, which addresses potential shortcomings and areas for improvement.

3.1 Setup and Protocol

The literature review process for this thesis is deliberately conducted through a manual, exploratory search strategy. This approach is chosen to allow a more organic, reflective mapping of relevant themes and concepts, following a thought-driven process rather than relying solely on review software. By tracing literature through logic and evolving connections, the review aims to ensure that key dimensions – particularly at the intersection of data science, infrastructure benchmarking, and ethics – are not prematurely filtered out by algorithmic selection mechanisms. This decision also aligns with the broader methodological philosophy of this thesis: that human judgment plays a critical role in guiding complex decision systems.

The structure of the review is informed by the SALSA framework (Search, Appraisal, Synthesis, Analysis) (Grant et al., 2009), which supports a transparent and staged engagement with academic literature. The first three stages are reflected in this section, while the Analysis is carried forward into later chapters where the model design is developed and critically applied. Inspiration is also drawn from previous structured reviews conducted at Delft University of Technology, including a notable study on algorithmic decision-making (Buszydlik, 2024), which blended active learning tools with citation tracking. Although this review adapts its methodology to a more interpretive mode, it shares the same goal: building a robust and contextually grounded foundation for the research that follows.

3.1.1 Search

To ensure the literature review is grounded in the full scope of the thesis objective, the search strategy is carefully shaped by its interdisciplinary nature. Three dimensions – **Technical**, **Water**, and **Ethical** – emerge as core thematic pillars through an iterative reflection on the main research question at hand, the practical demands of the World Bank assignment, and early engagement with relevant academic and policy literature. The search is therefore structured to capture contributions across both computer science and water sector domains, with particular attention to frameworks, challenges, and methodologies that operate at their intersection.

The search used Google Scholar[28] due to its accessibility and broad coverage of academic literature, and specialised databases such as ACM Digital Library[29], IEEE Xplore[30], and SCOPUS[31]. This broad scope aims to capture a more extensive range of publications.

The **technical dimension** of this literature review is guided by a central sub-question: *What technical approaches have been proposed for automating data validation and anomaly detection, and how do they account for uncertainty, scale, and limited ground truth?* This question emerges directly from the gaps identified in Chapter 2[32], where the current NewIBNET system is shown to rely on a single expert reviewer conducting manual checks across hundreds of submitted entries each year. The process is not only time-consuming, but also challenged by limited transparency, inconsistent comparator logic, and the absence of strong ground truth data. These observations form the initial hypothesis that anomaly detection methods could support a more efficient and interpretable validation pipeline.

Accordingly, the literature search focuses on identifying exist-

[28] **Google Scholar:** https://scholar.google.com/

[29] **ACM Digital Library:** https://dl.acm.org/

[30] **IEEE Xplore:** https://ieeexplore.ieee.org/Xplore/home.jsp

[31] **SCOPUS:** https://www.scopus.com

[32] *"...the process moves to the* ***Data Review*** *stage, where a single reviewer within the NewIBNET team examines the raw data for over 250+ utilities..."* in Chapter 2

ing methodologies for detecting inconsistencies within benchmarking data. This includes statistical outlier detection[33], machine learning techniques[34], and consistency checks[35], all of which serve as potential foundations for building anomaly detection mechanisms adapted to NewIBNET's context.

The search also considered various decision algorithms[36], particularly those designed to score and prioritise flagged entries, enabling a reviewer to focus attention where it is most needed. Since the NewIBNET system lacks labelled data and consistent flagging histories, special attention is paid to models that function in data-sparse environments or that incorporate uncertainty into their assessments.

The **water dimension** of this literature review is guided by the following sub-question: *How have benchmarking practices in the water and infrastructure sectors evolved in response to challenges of participation, comparability, and data quality?* The aim of this dimension is to position the thesis within the broader context of water utility benchmarking. The review examines research focused on IBNET-based reporting, as well as literature analysing other benchmarking platforms operating at regional, national, and global levels. These sources provide valuable insights into how data is collected and used within benchmarking efforts, and what institutional or technical strategies have been employed to manage diverse contexts. By synthesising these strands, the thesis seeks to explore a validation approach that is informed not only by data, but by a deeper understanding of the environment in which such data is produced and submitted.

The **ethical dimension** of the literature review is guided by the following sub-question: *What ethical tensions arise in automating public-sector data review, particularly around framing, transparency, and institutional trust?* This inquiry reflects a central concern of the thesis: how to form validation mechanisms that do not simply perform well, but also uphold the values and expectations embedded in public governance. The review focuses on literature that explores the ethical stakes of algorithmic decision-making in institutional settings, with particular attention to questions of bias, interpretability, and the role of human oversight.

Rather than treating fairness or accountability as add-ons to a technical model, these works position them as foundational design principles – influencing everything from system architecture to how outputs are communicated and used. This perspective is critical for the thesis, which operates in a domain where data-driven

[33] **Statistical Outlier Detection:** The process of identifying data points that significantly deviate from the majority of a dataset, potentially indicating errors or novel insights (Austin, 2004).

[34] **Machine Learning Techniques:** Computational methods that enable systems to learn from data and improve performance over time without being explicitly programmed (Mitchell, 1997).

[35] **Consistency Checks:** Procedures used to ensure data integrity by verifying that data entries conform to predefined rules or constraints (Batini et al., 2020).

[36] **Decision Algorithms:** Systematic procedures or formulas designed to aid in making choices by processing data and evaluating possible outcomes (Russel et al., 2006).

insights may influence real-world policy, funding, and utility performance narratives. As such, the ethical dimension not only informs model constraints, but shapes how trust and legitimacy are conceptualised within the validation process itself.

3.1.2 Appraisal

As outlined in the previous subsection, the literature review is constructed through a manual, SALSA-inspired search strategy, guided by the thematic dimensions of the thesis. Over 100 academic sources are selected through targeted keyword queries and citation tracking, with a focus on maintaining conceptual relevance to the central thesis objective.

Methodological rigour is evaluated based on the clarity and transparency of each study's research design, the presence of empirical grounding or applied frameworks, and peer-reviewed credibility. Studies offering generic technical overviews without contextual application to data, or those addressing peripheral topics without a clear link to benchmarking or decision-support systems, are excluded. Likewise, sources that lack specificity in their handling of data quality are filtered out, unless they contributed to conceptual discussions around reliability or comparability in public-sector data environments.

The final body of literature represents a diverse set of domains, contributing to a multi-layered understanding of the challenges this thesis seeks to address. To support structure and coherence, the reviewed work is organised into **seven thematic categories**.

3.1.3 Synthesis

To synthesise the findings, a thematic content analysis[37] is conducted. The process began with the three guiding sub-questions, which served as the initial lens for organising the literature along three dimensions: **Technical**, **Water**, and **Ethical**.

From an initial sweep of the literature, recurring themes were identified within each dimension. These themes provided a structured pipeline for subsequent review: once identified, they guided further targeted searches to ensure sufficient depth and coverage. In this way, the synthesis balances breadth with focus, while remaining transparent about the analytical lens applied.

The final framework consisted of seven themes, grouped per dimension:

[37] Origins of thematic content analysis often traced to Boyatzis, 1998, who formalised it as a systematic coding and theme-development method. The themes used in this thesis are specified in Section 3.2.

- **Technical:** (1) Contributions to Automated Data Validation and Benchmarking; (2) Criteria for Data Validation and Quality Assurance; (3) Decision Algorithms and Weighting Mechanisms; (4) Handling Data without Historical Labels.
- **Water:** (5) Use of Benchmarking Data and Related Datasets.
- **Ethical:** (6)Integration of Human Feedback in Automated Systems; (7) Ethical Considerations and Transparency.

Throughout the synthesis, attention is paid only to explicit claims and methods as articulated by the authors, avoiding interpretative overreach to maintain objectivity in thematic grouping.

> **Summary:** This section outlines the literature review approach, which was conducted manually using the SALSA framework to allow for a reflective, interdisciplinary mapping of relevant research. Guided by three core dimensions – Technical, Water, and Ethical – the review deliberately balanced anomaly detection methods with domain-specific benchmarking insights and ethical design principles. Sources were appraised for relevance and methodological rigour, and synthesised into seven thematic categories that directly inform the thesis' model architecture.

3.2 Thematic Content Analysis

Drawing on a diverse yet interrelated body of literature, ranging from machine learning, water systems engineering, ethics, and sociotechnical systems design, this analysis allows for the unfolding of seven key thematic areas.

3.2.1 Contributions to Automated Data Validation and Benchmarking

Anomaly detection emerges as one of the most frequently addressed topics in the reviewed literature, with 10 out of the surveyed sources discussing methods to identify irregularities in water data or similar domains. Techniques range from traditional statistical and clustering methods[38] (Chandola et al., 2009; Ahmed et al., 2016; Bhuyan et al., 2013; Tukey, 1977) to machine learning and deep learning-based models (Pang et al., 2021; Dogo et al., 2019; Kanyama et al., 2024; Nofal et al., 2021; Candelieri, 2017), including time-series decomposition[39] and domain-informed classification (Wu et al., 2021).

[38] **Clustering Methods:** Unsupervised techniques that group similar data points together to detect patterns or anomalies without predefined labels.

[39] **Time-Series Decomposition:** Breaks down time-series data into trend, seasonal, and residual components to better understand patterns over time (Kim et al., 2024; Lim et al., 2023).

These studies emphasise the importance of detecting inconsistencies in both real-time and historical data environments, with several highlighting scalability and adaptability as key advantages. The collective insights from this body of work form a foundational basis for exploring a data-driven, automated flagging system in benchmarking frameworks like NewIBNET (Raciti et al., 2012).

In parallel, data validation has emerged as a critical pillar supporting automated benchmarking, with a growing body of literature proposing innovative approaches for improving data reliability. Several studies introduce systems that incorporate domain knowledge, metadata, and structural constraints to detect errors before model training or data integration (Bachinger et al., 2024; Shankar et al., 2023). For instance, Bachinger et al., 2024 emphasises the use of shape constraints and expert knowledge to guide validation, while Shankar et al., 2023 proposes GATE[40], a system that summarises partitions to uncover corruptions over time. Other notable contributions include Y. Liu et al., 2025's dual-layer metadata validation[41], Peleska et al., 2021's logical rule-based framework for interlocking systems, and Song et al., 2021's Auto-Validate, which offers a rule-free, unsupervised approach[42] tailored for large-scale data lakes[43]. Collectively, these works advance the field by showcasing scalable, interpretable validation frameworks that could be adapted to sector-specific databases such as NewIBNET, where both reliability and automation are essential.

The theme of benchmarking surfaces across various domains, each contributing distinct perspectives on how performance measurement frameworks can drive operational improvements. Several studies explore benchmarking beyond water systems, including website governance (Misra et al., 2024), processor energy efficiency (Drávai et al., 2025), and AI model performance (Chitty-Vankata et al., 2025), offering transferable methodologies and assessment structures. Of particular relevance are efforts focused on utility-scale benchmarking for energy and water use in buildings and retail sectors (Dudani et al., 2022; Senanayake et al., 2012), as well as Mauro et al., 2023's dataset-driven benchmark for water resources monitoring using deep learning. While Surprise Benchmarking (Benson et al., 2024) introduces a novel idea of stress-testing systems with unpredictable workloads, its emphasis on resilience and adaptability offers useful parallels to how benchmarking systems like NewIBNET might evolve to accommodate diverse utility contexts and real-world uncertainty.

Performance evaluation is another recurring theme across the

[40] **GATE:** High-precision data validation system that detects corrupted data partitions by monitoring aggregate statistics over clustered feature groups (Shankar et al., 2023).

[41] **Dual-layer Metadata Validation:** Uses both front-end and back-end checks to ensure data quality by using metadata at multiple points in the data pipeline (Y. Liu et al., 2025).

[42] **Unsupervised Approach:** Refers to a machine learning method that identifies patterns or anomalies in data without requiring labelled examples or prior training outcomes.

[43] **Data Lakes:** Centralised repositories that store large volumes of raw, unstructured, semi-structured, or structured data in its native format for later processing and analysis.

literature, particularly within the water sector, where robust assessment models are essential for identifying inefficiencies and guiding strategic improvements. A number of studies propose structured evaluation frameworks such as the Super-Efficiency DEA (SE-DEA)[44] model (Yang et al., 2011) and fuzzy comprehensive evaluation approaches (Meng et al., 2021), both of which accommodate complex, multidimensional performance indicators. In addition to this, multicriteria decision-making (MCDM)[45] frameworks are used to assess sustainability and operational effectiveness in urban water supply systems (Wibowo et al., 2018; Wibowo et al., 2017), while case studies on non-revenue water reduction strategies provide insight into practical performance metrics for utility optimisation (Silva et al., 2023). These contributions not only highlight the importance of rigorous performance tracking but also emphasise the value of integrating quantitative analysis with contextual understanding.

Finally, classification-based approaches[46] appear across multiple reviewed studies, offering tailored solutions to automated categorisation tasks that mirror the anomaly detection and validation objectives in this thesis. Approaches range from binary classification for water potability prediction using models like logistic regression and support vector machines[47] (Castillo et al., 2024), to semi-supervised multi-class classifiers for real-time streaming data (C. Liu et al., 2025). Text and link classification models demonstrate applications in knowledge organisation (N. Liu et al., 2007), while supervised learning frameworks for crypto asset classification highlighted the role of transparent and interpretable categorisation. These methodologies not only support technical precision but also serve as inspiration for developing scalable and explainable flagging strategies within the context of NewIBNET's performance monitoring.

3.2.2 Criteria for Data Validation and Quality Assurance

A core pillar of data validation lies in the assurance of data quality – particularly in sparse heterogeneous systems like NewIBNET. Literature on this theme consistently highlights the need for structured, intelligent frameworks capable of managing complex datasets (Taleb et al., 2018; Qi et al., 2024; Kang et al., 2017). The works of Taleb et al., 2018 and Kang et al., 2017 offer valuable overviews of quality assessment in unstructured and predictive contexts, while Qi et al., 2024 emphasise the role of AI in building decision support systems for water quality. Sector-specific contributions such as the

[44] **Super-Efficiency Data Envelopment Analysis:** An extension of the traditional DEA model that allows for ranking efficient decision-making units by measuring their performance beyond the efficient frontier (Yang et al., 2011).

[45] **Multicriteria Decision-Making (MCDM):** Set of analytical methods used to evaluate and prioritise multiple conflicting criteria in decision-making processes (Wibowo et al., 2018).

[46] **Classification-based Approaches:** Use machine learning models to assign data points into predefined categories or classes, often applied in anomaly detection to label data as ”normal” or ”anomalous”.

[47] **Support Vector Machines (SVMs):** Supervised learning models used for classification and anomaly detection by finding the optimal boundary that separates different classes in a dataset.

work of Yu et al., 2023 demonstrate how weighted index methods like WQI[48] can provide quantifiable quality scores, useful for systematised flagging. Broader institutional efforts, like those from the World Bank, 2022 and the European Data Quality Guidelines[49], introduce best practices for data governance, verification, and lifecycle management, offering transferable strategies for NewIBNET. These sources form a foundation for designing robust data quality checks within an automated validation model, enabling both technical precision and stakeholder trust.

Complementing the broader discussions on data quality, a smaller yet significant strand of literature focuses on enhancing data consistency, particularly when integrating or validating inputs from disparate sources. Abián et al., 2019 introduce contemporary constraints as a dynamic method to flag temporal inconsistencies within datasets, while Deng et al., 2022 propose a novel consistency index to assess multi-measurement agreement. Huang et al., 2019 extend this work to the multi-source domain through matching dependencies, providing a structured approach to resolve contradictions in integrated data systems.

A central challenge within benchmarking datasets – especially those with voluntary reporting and uneven coverage, such as NewIBNET – is how to address missing data without introducing noise or bias. The reviewed studies highlight several strategies for managing incomplete datasets in ways that vary significantly in robustness and interpretability.

Bicego et al., 2024 introduce a Random Forest-based[50] distance metric that operates without imputation[51], allowing for similarity assessment even when data entries are partially missing. This nonparametric method shows promise in preserving structural relationships, making it potentially suitable for detecting outliers in sparse or noisy datasets. In contrast, X. Wang et al., 2023 argue for preserving the distribution of missingness when generating synthetic data[52] – a strategy more aligned with data augmentation than validation, and potentially less robust in flagging subtle inconsistencies. Meanwhile, Clifton et al., 2022 propose a privacy-preserving imputation approach using differentially private[53] k-NN[54] models. While valuable for protecting sensitive records, the technique may introduce smoothing that risks masking meaningful deviations – a trade-off to consider in small, skewed datasets.

An empirical benchmark by Miao et al., 2024 compares state-of-the-art imputation methods across various scenarios, offering practical insights into their performance under different data con-

[48] **Water Quality Index (WQI):** Composite metric that aggregates multiple water quality parameters into a single score to assess overall water quality (Yu et al., 2023).

[49] **European Data Quality Guidelines:** Read more: https://op.europa.eu/webpub/op/data-quality-guidelines/en/

[50] **Random Forest:** Ensemble learning method that constructs multiple decision trees to improve classification accuracy and handle missing data effectively (Bicego et al., 2024).

[51] **Imputation Algorithms:** Methods used to estimate and replace missing data within a dataset to enable complete data analysis.

[52] **Synthetic Data:** Refers to artificially generated data that mimics real datasets, often used to preserve privacy or augment limited data.

[53] **Differentially Private:** A formal privacy technique that ensures individual data entries cannot be re-identified, even from aggregated results.

[54] **k-Nearest Neighbours (k-NN):** A machine learning algorithm that classifies or imputes data based on the closest k similar data points in the dataset.

ditions. Finally, domain-specific applications such as water quality forecasting (Dong et al., 2024) and classification systems (Li et al., 2023) demonstrate that imputation strategies can significantly affect downstream accuracy, further highlighting the importance of context-specific method selection.

Across these contributions, it becomes clear that no single method universally fits sparse benchmarking datasets; instead, trade-offs between completeness, transparency, and anomaly preservation must be carefully weighed in the design of a validation system.

Although terms like *data reliability* or *deviation analysis* rarely appear explicitly in the reviewed literature, many of the underlying concerns – such as robustness under uncertainty and the risk of masking meaningful anomalies – are addressed through adjacent concepts in data consistency, quality assurance, and missing data handling.

3.2.3 Decision Algorithms and Weighting Mechanisms

In designing intelligent systems for complex socio-environmental challenges, robust decision algorithms and weighting mechanisms are essential – not only for accurate anomaly detection but also for ensuring that automated processes remain adaptable, transparent, and context-aware. Across the literature, a clear convergence emerges: decision support systems must balance algorithmic precision with flexibility to accommodate multi-dimensional, often uncertain, data environments. From hybrid machine learning models that blend decision tree logic with probabilistic inference to boost performance under data variability (Hall, 2007; Jia, 2022), to AI-augmented operational planning tools that integrate real-time feedback loops for water resource optimisation (Xian et al., 2024; Jalal et al., 2020), the spectrum of innovation reveals a shared emphasis on contextualisation, weighting, and transparency. Several approaches foreground the importance of domain-specific heuristics, such as similarity-based reasoning (Zeng et al., 2012) or multi-criteria evaluation strategies (Caylor et al., 2020), to guide complex choices in resource-constrained and dynamic environments. These systems move beyond static rules to support iterative, learning-based governance, where decisions on flagging inconsistencies must reflect both statistical outliers and normative benchmarks. Underpinning many of these models is a growing reliance on fuzzy or linguistic decision-making paradigms (Herrera-Viedma et al., 2020), which prove particularly effective in navigating ambiguity – whether stemming from incomplete data or competing stakeholder values. These contributions show that a well-designed decision algorithm is not simply an optimisation engine but a dynamic interface between data, governance, and judgment.

Scoring systems have emerged as mechanisms for translating complex, multidimensional inputs into interpretable outputs that inform action. Recent advancements – from fuzzy logic models integrating expert-derived weights (Kahla et al., 2025), to transformer-based architectures that isolate temporal deviations in real-time data streams (Kim et al., 2024) – demonstrate the growing precision and adaptability of score-based approaches. Whether optimising detection efficiency in large-scale infrastructures (Chang, 2024), flagging inconsistencies in categorical datasets via recommender logic (Belgacem et al., 2024), or distinguishing anomalies through refined score distribution modelling (Lim et al., 2023; Jiang et al., 2023), these methods collectively emphasise that scoring is no longer a static threshold but a dynamic, contextual process. This evolution aligns with the vision for a framework: a system that accounts for subtle deviations, scalable processing, and meaningful severity rankings – anchored not in rigid cutoffs, but in intelligent, evidence-weighted interpretation.

An increasingly sophisticated challenge is not only identifying whether an anomaly exists, but determining *how severe* it is – especially in contexts where nuanced gradations of concern can trigger vastly different operational responses. The literature signals a maturing shift toward integrating severity-oriented weighting mechanisms into algorithmic frameworks, thus refining the value and interpretability of flagged anomalies. Emerging models such as Fuzzy Weighted Principal Component Analysis (FWPCA)[55] (S. Wang et al., n.d.) and Bi-Bayesian Gaussian Mixture Models (Bi-BGMM)[56] (Bingöl et al., 2024) introduce layered sensitivity to contextual variables by assigning adaptive weights to both features and instances, capturing subtle, high-dimensional deviations that would otherwise be flattened in binary anomaly labelling. Complementarily, Score Distribution Discrimination[57] methods and severity-linked performance metrics (Hajirahimi et al., 2023; Yi et al., 2024) enhance evaluative accuracy by measuring alignment between predicted and actual severity levels, a crucial step when moving from detection to actionable insights. Particularly relevant is the classification of anomaly severity tiers in multi-class frameworks, as demonstrated in drone fault detection (Silalahi et al., 2024), which resonates with the thesis objective of developing a system that pri-

[55] **Fuzzy Weighted Principal Component Analysis (FWPCA):** A dimensionality reduction technique that integrates fuzzy logic and weighted PCA to enhance anomaly detection by effectively handling data uncertainty and emphasising significant features (S. Wang et al., n.d.).

[56] **Bi-Bayesian Gaussian Mixture Models (Bi-BGMM):** Dual-stage anomaly detection method that employs Bayesian Gaussian Mixture Models to perform bi-clustering, simultaneously identifying anomalous patterns across both data features and instances (Bingöl et al., 2024).

[57] **Score Distribution Discrimination:** Technique that evaluates and differentiates multiple anomaly score distributions to enhance the accuracy of anomaly detection systems (Yi et al., 2024).

oritises and stratifies anomalies based on their urgency and systemic impact.

Taken together, these methodologies advocate for moving beyond simple anomaly thresholds to develop weighted, interpretable, and severity-informed systems. While many of these models have been developed and tested in high-volume, high-density data environments, their conceptual relevance remains significant for this thesis. As data platforms such as NewIBNET continue to expand, understanding how severity-based logic can be layered into validation systems offers a forward-looking perspective – one that anticipates future scale while remaining grounded in the practical constraints of its current data regime.

Lastly, fuzzy logic has become known as a powerful decision-making paradigm in contexts characterised by uncertainty, ambiguity, and the need for interpretability – features that are acutely relevant to water utility data environments. Across multiple studies, fuzzy inference systems have demonstrated their ability to translate imprecise numerical data into actionable and linguistically meaningful classifications, whether through real-time water quality monitoring (Paul B. Bokingkito et al., 2018), evaluating filtration systems (Yumang et al., 2021), or enabling responsive water supply management in uncertain conditions (Sharma et al., 2012). The strength of fuzzy logic lies in its alignment with human reasoning – facilitating nuanced assessments such as classifying water as "Good" or "Poor," rather than relying on rigid thresholds. Recent developments in hybrid models, such as fuzzy neural networks (Lin et al., 2025; Zhu, 2009), extend this further by combining the transparency of fuzzy rules with the learning capacity of neural architectures – enabling robust classification even in noisy or non-linear datasets. In addition to this, the application of fuzzy logic in modelling human-machine interactions (Cui, 2024) and decision factor weighting (Anifa et al., 2024) highlights its flexibility across domains, particularly in interfacing human judgment with automated systems.

3.2.4 Handling Data Without Historical Labels

The challenge of performing historical consistency checks in the absence or limitation of past data has been addressed through a range of innovative applications of established techniques. Dai et al., 2021 propose similarity-based forecasting models to predict sales for products without historical records, demonstrating the viability of reference-based estimation strategies. Winona et al., 2020 apply Long Short-Term Memory (LSTM) networks[58] – a long-standing temporal modelling method – to generate short-term sea level predictions using minimal prior data. Similarly, Singhal et al., 2001 apply Principal Component Analysis (PCA)[59] in a novel framework to assess current values against latent historical structures, and Aubry, 2021 explores the broader capacity of deep learning to extract meaningful signals from fragmented legacy datasets.

These contributions highlight how classical mathematical techniques, when creatively applied to data-sparse scenarios, can offer valuable alternatives to traditional historical consistency checks.

3.2.5 Integration of Human Feedback in Automated Systems

The integration of human feedback into automated systems is a core theme across several reviewed studies, emphasising that sustainable system performance hinges not only on technical soundness but also on the active engagement of end users throughout the system's lifecycle. Socio-technical perspectives foregrounded by Baxter et al., 2011 and extended by Sommerville et al., 2007 underline the necessity of designing systems that account for layered human realities – technical, organisational, and social. In the context of international systems like NewIBNET, where users operate under diverse institutional and cultural settings, this perspective becomes not just relevant but essential. The principles from Human-in-the-Loop Machine Learning (HITL-ML)[60], as explored by Munro, 2021 and Kumar et al., 2024, offer tangible strategies – such as iterative annotation, feedback-driven refinement, and domain expert integration. These insights are further supported by Tong et al., 2009, whose framing of product innovation highlights how iterative co-design, shaped by real-time market and user response, drives long-term system trust and usability.

3.2.6 Ethical Considerations and Transparency

The ethical dimensions of algorithmic decision-making, particularly in the context of automated flagging systems like those envisioned for NewIBNET, must be addressed not as an afterthought but as an integral part of the design process. As emphasised by Waller et al., 2025, bias mitigation in binary classification systems cannot be reduced to technical fixes alone – it requires ethically sound, legally compliant, and contextually aware approaches from

[58] **Long Short-Term Memory (LSTM) Networks:** A type of recurrent neural network (RNN) first introduced by Hochreiter & Schmidhuber (1997), designed to model temporal sequences and capture long-range dependencies (Hochreiter et al., 1997, Winona et al., 2020).

[59] **Principal Component Analysis (PCA):** A statistical technique introduced by Pearson (1901) for reducing data dimensionality while preserving variance. See more: `http://www.stats.org.uk/pca/Pearson1901.pdf`

[60] **Human-in-the-Loop Machine Learning (HITL-ML):** Integrates human expertise into the training and refinement of machine learning models, allowing for active human involvement in data annotation, model validation, and decision-making processes to enhance model accuracy and reliability (Munro, 2021; Kumar et al., 2024).

the outset. Building on this, the ethical cycle framework of van de Poel et al., 2011 highlights the necessity of embedding values and stakeholder engagement throughout the lifecycle of technological development, rather than retrofitting ethical considerations post-implementation. Equally critical is the role of linguistic framing: McNealy, 2021 highlights how the language used to communicate ethics directly affects public trust and system legitimacy, reinforcing the idea that transparency is not merely a design feature but a narrative tool. Complementing this, Borghouts et al., 2024's study on COVID-19 vaccine messaging demonstrates how subtle variations in emotional tone and certainty can drastically alter user reception, particularly across ideological lines – insights that are vital when considering how anomalies or inconsistencies in utility data are flagged and presented to global reviewers. These sources argue persuasively that transparency, fairness, and cultural sensitivity must be interwoven into the core architecture of systems like NewIBNET to promote trust, reduce algorithmic harm, and support equitable utility governance across diverse socio-political landscapes.

3.2.7 Use of Benchmarking Data and Related Datasets

The literature presents IBNET as a foundational data source for evaluating utility performance on a global scale. In policy-driven research, IBNET is valued for enabling comparative political economy analysis and facilitating evidence-based decision-making, as demonstrated by Manghee et al., 2012. Other studies indicate IBNET's role as both a benchmarking tool and a data repository, while simultaneously identifying structural limitations. For instance, Andres et al., 2020 directly engage with IBNET's high degree of missing data, proposing a nested panel methodology to improve its analytical usability. Similarly, C. v. d. Berg et al., 2017 draw extensively from IBNET to assess African utility performance, positioning it as a transparency-enabling platform for governments and operators alike. However, Bhatt, 2024 offers a necessary critique, highlighting how IBNET's efficiency-centric indicators may obscure local equity concerns. Additional empirical studies employ IBNET to support investigations into tariff subsidies (Andrés et al., 2020), fraud detection frameworks (Detroz et al., 2017), and financial-service quality relationships (Tsagarakis, 2018), all of which signal its enduring relevance in both operational and developmental contexts. These applications reinforce a necessity for NewIBNET: to optimise current validation framework, ensuring that the system's data quality can meet the increasing demand for robust, contextualised benchmarking.

A search for academic literature referencing *NewIBNET* yielded no direct results, likely due to its recent launch and the time required for uptake in scholarly publications. However, given IBNET's proven value in utility benchmarking, NewIBNET holds considerable promise in continuing this legacy and in potentially addressing the data quality and structural shortcomings identified in the earlier system.

Beyond the widespread use of IBNET in global utility benchmarking, several regional and thematic platforms offer complementary approaches that can inform the evolution of NewIBNET. The AWWA Utility Benchmarking Program[61] in North America emphasises granular, high-frequency KPIs tailored to local regulatory and operational contexts, contrasting NewIBNET's broader, cross- country comparability. Similarly, the European Benchmarking Co-operation (EBC)[62] focuses on service quality and knowledge sharing within Europe, providing region-specific depth while lacking NewIBNET's global inclusivity. The Utility Benchmarking Program (UBP)[63] by IAWD, centred on the Danube region, exemplifies a subregional model where benchmarking is closely aligned with local environmental and institutional conditions. These platforms highlight the trade-offs between contextual specificity and global comparability. A detailed comparative analysis of these benchmarking platforms in relation to NewIBNET is provided in Appendix B.1.

Several academic contributions also propose tailored performance indicator systems, such as Ganjidoost et al., 2018's time-integrated benchmarking for pipelines, which shows the need for localised indicators that account for demographic and infrastructure variability. Others, like Burdescu et al., 2020's Caribbean utility benchmark, align with NewIBNET's goals but demonstrate how contextual adaptation – through frameworks like the Water Utility Turnaround model – can yield regionally actionable insights. These tools could offer valuable lessons for refining NewIBNET's structure, particularly regarding indicator standardisation, regional adaptability, and the balance between breadth and depth in utility performance comparison.

In addition to the structural comparisons above, the literature found also offers insights into the practical incentives and contextual constraints that influence utility participation in benchmarking platforms like NewIBNET. S. Berg, 2010 emphasise the role of

[61] **AWWA Utility Benchmarking Program:** Read more: `https://www.awwa.org/programs/benchmarking`

[62] **European Benchmarking Co-operation (EBC):** Read more: `https://www.waterbenchmark.org/`

[63] **Utility Benchmarking Program (UBP):** Read more: `https://www.iawd.at/eng/danube-toolbox/d-leap/programs/utility-benchmarking-program/`

donor pressure, performance-based funding, and institutional reputation as key motivators, particularly in low- and middle-income countries. They highlight how utilities are more likely to engage in benchmarking when results feed into national policy frameworks or funding eligibility. At the same time, reporting challenges such as limited technical capacity, differing interpretations of indicator definitions, and political sensitivities surrounding transparency introduce risks of partial or misreported data. These observations highlight the importance of not only automating validation processes, but designing them with a realistic understanding of the operational, financial, and institutional environments in which utilities operate. While such domain-specific constraints may not always be visible in purely technical literature, they form a necessary bridge between system design and real-world implementation – a gap that this thesis seeks to navigate.

3.3 Key Insights from the Literature Review

This section integrates insights from all interdisciplinary sources to reflect critically on the thematic landscape uncovered through the literature review. By weaving together methodological innovations, sectoral realities, and normative considerations, it positions the literature review at the intersection of data science, governance, and responsible system design.

3.3.1 Technical Dimension

> *What technical approaches have been proposed for automating data validation and anomaly detection, and how do they account for uncertainty, scale, and limited ground truth?*

The development of automated validation systems in the context of global water utility benchmarking requires integrating techniques that are not only statistically rigorous but also context-aware, scalable, and robust to uncertainty. Literature across domains such as anomaly detection and decision theory highlights a progression from traditional rule-based screening toward hybrid frameworks that combine statistical inference, structural logic, and interpretability.

While Section 3.2 reviewed individual approaches to anomaly detection, the synthesis here highlights a broader insight: anomalies are not fixed entities but context-dependent deviations from an evolving baseline of ”normal” behaviour (Chandola et al., 2009). Detecting such deviations necessitates a foundational layer of statistical profiling, where descriptive metrics – such as means, medians, and standard deviations – offer a first pass at distinguishing expected values from outliers (Ahmed et al., 2016; Bhuyan et al., 2013). Techniques like histogram visualisation further support the identification of skewed or multimodal distributions, helping to establish empirical baselines against which irregularities may be judged (Chandola et al., 2009). More broadly, such practices align with the principles of exploratory data analysis[64], where the combined use of descriptive statistics and visualisation serves as a general analytic strategy for uncovering structure, detecting anomalies, and suggesting underlying patterns in the data (Tukey, 1977).

However, statistical profiling alone cannot resolve structural inconsistencies or embedded entry errors. Here, the literature points to schema-based validation and rule-driven conditional logic as essential complements (Peleska et al., 2021). While often underutilised in benchmarking contexts, such checks remain critical for ensuring internal coherence, including plausibility thresholds and logical dependencies across fields. This aligns with early-stage anomaly detection practices in network security and financial auditing, where metadata-driven assumptions – such as declared input types or structural constraints – are used to screen for irregular submissions before deeper modelling is applied.

Where data gaps persist, imputation strategies become central. Although statistical methods such as median imputation are widely used in resource-constrained settings (Miao et al., 2024), literature in privacy-preserving systems points to the potential of k-nearest neighbour (K-NN) approaches (Clifton et al., 2022), particularly in settings with partial correlation structures. Given the variability in data availability across utilities, the thesis could test both approaches to evaluate their impact on downstream validation accuracy, while recognising the importance of maintaining traceability and interpretability in imputed values.

Building on this foundation, the next layer of experimentation engages with comparator-based anomaly modelling, where the challenge shifts from detecting irregularities in isolation to doing so relative to meaningful peer groups. The literature on benchmarking systems, including critical assessments of IBNET, highlights the importance of embedding utility metadata[65] such as region, income level, and population served into comparative frameworks

[64] **Exploratory Data Analysis (EDA):** The process of analysing datasets to summarise their main characteristics, often using visual methods (Tukey, 1977).

[65] As highlighted in Chapter 2, reliance on global averages alone may overlook important contextual differences between utilities.

(Manghee et al., 2012; Bhatt, 2024; Tsagarakis, 2018). These dimensions reflect not only service context but also institutional and financial variation, making them key axes along which deviation should be measured.

Traditional statistical methods like z-score calculation, already used manually in NewIBNET, remain central to this task, standardising deviations relative to comparator group distributions. However, the literature also suggests experimenting with composite z-score profiling and weighted normalisation, where the influence of each comparator group is adjusted based on statistical correlation or policy relevance (Yu et al., 2023). This motivates a more layered understanding of what constitutes a significant anomaly: one that may stand out across multiple benchmarks, or one that is extreme relative to a particularly relevant dimension (e.g., comparator group peers).

To transform these detection outputs into decision support tools, the literature introduces methods from anomaly scoring and severity prioritisation. Multi-criteria decision-making frameworks and static/dynamic weighting models offer templates for building composite severity scores (Kim et al., 2024; Hajirahimi et al., 2023). This thesis could adapt those insights to the benchmarking context by combining z-score deviation with fixed comparator weights – a practical choice justified both by literature and the need for transparency in public-sector settings. Complementing this, severity tiers and thresholds, inspired by anomaly event classification in time-series systems (Wu et al., 2021), allow for a more nuanced, ranked interpretation of flagged outputs, making the system more actionable for reviewers.

Final validation strategies are informed by research into socio-technical system design. Literature emphasising human-in-the-loop models and dependable system validation (Baxter et al., 2011; Sommerville et al., 2007) stresses the importance of aligning automated decisions with expert expectations. This justifies the inclusion of expert surveys, manual benchmark comparisons, and similar case-based evaluations in the experimental framework, allowing the proposed system to be tested not only for statistical coherence, but also for institutional credibility and relevance.

Much of the literature reviewed in these themes proposed advanced anomaly detection and machine learning techniques, often tailored to contexts with abundant labelled data and strong computational resources. While such methods are powerful, they are ill-suited to the sparse, heterogeneous, and politically sensitive environment of global water utility benchmarking, where transparency and interpretability are paramount. Rather than selecting a single model or method, the technical strand of this thesis assembles an integrated pipeline – grounded in literature – that supports scalable, context-sensitive, and trustworthy anomaly detection within the realities of global public infrastructure.

> **Summary:** This section synthesises technical literature on statistical profiling, rule-based validation, comparator-based modelling, and severity scoring to inform the design of an automated anomaly detection system for water utility benchmarking. It highlights the importance of balancing scalability, and uncertainty in sparse, heterogeneous datasets like NewIBNET. The selected focus methods can be integrated into a layered pipeline that reflects both empirical evidence and practical constraints of global public-sector platforms.

3.3.2 Water Dimension

> *How have benchmarking practices in the water and infrastructure sectors evolved in response to challenges of participation, comparability, and data quality?*

A recurring insight from the literature engaging directly with IBNET is the central role benchmarking plays in both diagnosing sectoral inefficiencies and enabling informed policy interventions across diverse governance contexts. Studies such as Manghee et al., 2012 emphasise IBNET's capacity to facilitate political economy analysis, not just as a technical platform but as an evidence-based tool for navigating reform. Others, like Andrés et al., 2020, highlight IBNET's ability to support financial transparency, particularly in subsidy evaluation and tariff structuring, illustrating how robust performance indicators can underpin equity-driven water governance. However, several sources also indicate systemic limitations that carry methodological relevance – for instance, high volumes of missing data (Andres et al., 2020) and an overemphasis on quantifiable efficiency that may unintentionally obscure deeper structural inequalities (Bhatt, 2024). These observations inform the thesis in two significant ways. First, they signal the need to construct a flagging system that is not only technically rigorous, but also capable of contextual prioritisation and adaptive equity-aware adjustments. Second, comparative platforms such as AWWA, UBP, and EBC demonstrate the value of regional granularity and user-centred indicator design. These systems show that benchmarks become more actionable when they align with the operational, regulatory, and environmental conditions of specific regions. As such, the methodology will consider how NewIBNET can strike a balance between global standardisation and local relevance, supporting both cross-country comparisons and context-specific flagging mechanisms that respond to actual utility needs. In addition, the upstream and downstream processes embedded in these platforms offer valuable architectural and procedural insights. These pipeline components will inform the broader design logic of this thesis, ensuring that the validation framework not only detects anomalies but integrates meaningfully into the full lifecycle of benchmarking and utility engagement.

Despite IBNET's wide presence in literature, no current studies explicitly mention or evaluate the recently launched NewIBNET system, reflecting a noticeable absence in the academic landscape. This gap highlights a need for research that not only investigates NewIBNET's potential improvements but also explores how its updated benchmarking logic and usability can address longstanding issues such as missing data, limited context sensitivity, and inadequate stakeholder alignment.

> **Summary:** Benchmarking practices in the water and infrastructure sectors have evolved from purely performance-oriented comparisons toward more participatory, context-sensitive frameworks that account for equity, data quality, and regional governance realities. This thesis builds on that evolution by examining how systems can operationalise these shifts, addressing gaps in comparability, participation, and data reliability through a redesigned, stakeholder-informed validation approach.

3.3.3 Ethical Dimension

> *What ethical tensions arise in automating public-sector data review, particularly around framing, transparency, and institutional trust?*

Within the literature there is a clear recognition that automation – especially in systems like NewIBNET – cannot exist in isolation from the human contexts in which it operates. The socio-technical lens articulated by Baxter et al., 2011 and further refined by Sommerville et al., 2007 emphasises that data validation systems must be co-designed with the users they intend to serve. In the context of global benchmarking, where regional disparities and political sensitivities influence reporting behaviour, anomaly flagging must be not only statistically accurate but also institutionally trustworthy and context-aware. This is reinforced by Munro, 2021's advocacy for human-in-the-loop machine learning models, which prioritise expert input in ambiguous cases through active learning, iterative annotation, and domain-informed validation checkpoints – practices that can significantly improve the relevance and acceptance of flags generated within NewIBNET. The integration of structured reviewer feedback loops, as suggested by Kumar et al., 2024, is particularly valuable in settings characterised by data uncertainty and variation in expertise across utilities. Tong et al., 2009 extend this idea into design management, proposing that anomaly detection systems be treated as iterative products – requiring continuous engagement with stakeholders, alignment between technical and functional teams, and the flexibility to adapt based on evolving user needs. These works show a critical shift in perspective: that automated data validation is not a one-off implementation, but a dynamic, dialogic process where algorithms and human experts co-create meaningful, actionable outputs.

Despite extensive theoretical work on human-in-the-loop systems and socio-technical design, current research lacks application to public-sector digital infrastructures, particularly in globally scaled benchmarking contexts such as NewIBNET. This thesis could bridge that gap by embedding feedback-driven anomaly validation mechanisms into a real-world, institutional water utility system, translating abstract design and ethical principles into concrete, governance-relevant functionality.

A core thread emerging from the ethical literature is the imperative to embed transparency, fairness, and contextual sensitivity not as post hoc concerns but as foundational design principles in any automated decision-making system. In the context of an anomaly detection system, the framing and communication of flagged anomalies must be carefully calibrated to ensure neutrality, especially when engaging with diverse utility operators across geopolitical and cultural contexts (McNealy, 2021; Borghouts et al., 2024). Linguistic choices can significantly influence the degree of reviewer trust and engagement, aligning with findings from studies on politically sensitive messaging (Borghouts et al., 2024). Beyond language, legal and procedural transparency must underpin the system's structure, with bias mitigation approaches tailored not only for statistical fairness but also to align with frameworks like GDPR[66] and sector-specific norms (Waller et al., 2025). The ethical cycle, as proposed by van de Poel et al., 2011, offers a methodological anchor for this process: iteratively involving stakeholders, reassessing value trade-offs, and explicitly surfacing the societal implications of system outputs.

These insights shape the thesis's design priorities by embedding human dignity and institutional accountability into the validation logic, positioning automation not as a replacement for expert judgement, but as a partner in enabling equitable, transparent decision-making within global benchmarking systems like NewIBNET.

> **Summary:** Ethical tensions in automating public-sector data review arise from the need to balance algorithmic efficiency with institutional trust, particularly in how anomalies are framed and communicated to diverse stakeholders. Ensuring transparency, linguistic neutrality, and human-in-the-loop mechanisms is essential to prevent misinterpretation, build reviewer confidence, and uphold fairness in politically and culturally sensitive benchmarking contexts.

3.4 Limitations of the Literature Review

While the literature review provides a multidimensional understanding of automated benchmarking and anomaly detection in the context of water utility data, it is not without limitations. First, the inclusion criteria primarily focused on peer-reviewed academic

[66] **General Data Protection Regulation (GDPR):** A legal framework that governs how personal data is collected, processed, and stored within the European Union. Read more: https://gdpr-info.eu/

publications and institutional reports in English, which may have excluded relevant grey literature, practitioner insights, or region-specific case studies published in other languages. Additionally, while the SALSA framework supported a structured synthesis across technical, sectoral, and ethical dimensions, the review leaned more heavily toward methodological depth in the technical domain. This is partly due to the richer availability of literature on anomaly detection techniques compared to governance-oriented or utility- specific studies, which may have constrained the granularity of insights into institutional incentives and behavioural drivers behind benchmarking participation. As the review and synthesis is conducted by a single researcher, the analysis may also reflect individual blind spots or biases despite efforts to apply the framework systematically.

Moreover, although efforts are made to balance breadth and specificity, some selected models are designed for large, high- frequency datasets and may have limited transferability to sparse, heterogeneous datasets like those in NewIBNET. While their inclusion helped shape a forward-looking design space, the applicability of such models must be treated with caution and tested through practical adaptation. Finally, given the recency of NewIBNET's launch, no scholarly publications are found evaluating its architecture or implementation, limiting direct literature-based assessment of the system under study. This gap reinforces the importance of this thesis in contributing original research grounded in both conceptual synthesis and practical system engagement.

No automated software tools are used to conduct the literature review. While the exclusion of tools such as machine-assisted systematic review platforms is partly due to the thesis's alignment with human-in-the-loop principles, it is acknowledged that such tools could have been adapted to fit this methodology without undermining reviewer involvement. In addition to this, the manual approach allowed for more deliberate ethical discernment, especially in avoiding over-reliance on citation frequency as a proxy for influence – a practice that risks reinforcing structural biases in academic publishing, particularly when the subject matter intersects with developing country contexts. While automation could have enhanced coverage and consistency, its careful and ethically attuned integration remains a promising avenue for future iterations of similar research.

4 Methodology

This chapter outlines the thesis approach taken to explore automated data-driven validation. It presents the guiding research questions in Section 4.1, a design and methodological framework in Section 4.2, and a risk analysis in Section 4.3 – together forming the foundation for the system's development, validation, and ethical framing.

4.1 Research Questions

At the core of this research lies the following central **Research Question** (**RQ**):

> *How can data-driven mathematical models enhance validation and benchmarking of water utility indicators while ensuring reliability, decision-making integrity, and ethical transparency?*

This question reflects the dual ambition of the thesis: to introduce technical innovation in anomaly detection and indicator validation, and to do so in a way that aligns with the ethical, institutional, and user-centred realities explored in Chapters 2 and 3. The literature review established not only the technical opportunities but also the sector-specific constraints and ethical imperatives involved in public benchmarking systems.

To operationalise this objective, the thesis is structured around five core components. These components – drawn directly from the gaps and priorities identified in the literature – serve as thematic anchors, each accompanied by a guiding sub-question:

4.1.1 Data Preparation & Structural Validation

RQ1 *How can statistical profiling and rule-based logical checks be used to detect data quality issues and prepare water utility indicator data for reliable anomaly detection?*

This phase focuses on preparing the raw utility data for downstream anomaly detection. As discussed in Chapters 3, the literature highlights that anomalies are not fixed entities but context-dependent deviations, requiring a foundational layer of statistical profiling to establish baselines. Building on this, descriptive metrics and histogram visualisation are applied to surface outliers and distribution patterns (Chandola et al., 2009; Ahmed et al., 2016; Bhuyan et al., 2013; Wu et al., 2021). These are complemented by conditional logic and schema-driven rules derived from NewIBNET's structure and metadata, reflecting the importance of rule-based coherence checks (Raciti et al., 2012; Y. Liu et al., 2025). Where data gaps are detected, the approach will test both K-nearest neighbour imputation (Clifton et al., 2022) and median imputation as a simpler statistical method (Miao et al., 2024). These choices ensure that design decisions are not ad hoc but grounded in the thematic insights developed earlier, while being adapted to the specific constraints of the NewIBNET dataset.

4.1.2 Context-Aware Anomaly Modelling Frameworks

RQ2 *How can utility metadata and comparator-based modelling support the detection of deviations in performance indicators across diverse water utilities?*

Building on the cleaned dataset, the second phase introduces comparator-based modelling to move beyond universal thresholds and instead detect anomalies relative to meaningful peer groups. This is inspired by benchmarking research advocating for contextual validation frameworks and more nuanced comparator segmentation (Manghee et al., 2012; Bhatt, 2024; Tsagarakis, 2018). New comparator groups based on region and income level are analysed and extended. The chapter also looks into current z-score deviation calculations based on the current comparators.

4.1.3 Severity Scoring & Decision Framework

RQ3 *Which severity scoring methodologies can be investigated to best translate statistical deviations into a prioritisation of anomalies?*

Once statistical deviations are identified, the next step is to determine which anomalies should be prioritised for review. Building on decision science principles and insights from anomaly scoring literature (Kim et al., 2024; Hajirahimi et al., 2023), this phase investigates multiple severity scoring methodologies to translate deviation measures into a prioritisation of anomalies. The analysis compares fixed-weight schemes, variance-driven score calibration (Yu et al., 2023), and tiered severity classifications inspired by established anomaly categorisation frameworks (Wu et al., 2021).

4.1.4 Technical Validation & Evaluation

RQ4 *To what extent does the proposed anomaly flagging system perform reliably, and align with expert validation and benchmarking expectations?*

To understand how the system aligns with institutional needs and expert expectations, a final technical evaluation is crucial. This includes an expert survey and performance feedback loop, grounded in socio-technical system literature that emphasises human-centred AI and trustworthy automation (Baxter et al., 2011; Sommerville et al., 2007). Performance is further assessed through a targeted case study, and internal test scenarios, providing a multidimensional understanding of system reliability and interpretability in practice.

4.1.5 Ethical, Political, and Sectoral Considerations

RQ5 *What ethical and institutional implications arise from implementing an automated anomaly detection framework in the context of global water utility benchmarking?*

Beyond performance, the thesis closes with a reflective analysis of the ethical and institutional implications of automating decision support in a public-sector benchmarking system. Drawing from literature on bias mitigation, linguistic framing, and trust in automation (Waller et al., 2025; McNealy, 2021; Borghouts et al., 2024; Munro, 2021), the framework is evaluated for its alignment with fairness, transparency, and institutional legitimacy. These considerations inform not only the survey design and system language, but also broader reflections on the role of automated flagging in global development contexts.

Together, these research areas provide a structured lens through which the thesis explores the intersection of automation, benchmarking integrity, and ethical accountability, bridging technical innovation with real-world applicability and stakeholder trust.

> **Summary:** 5 sub-questions were formulated under five pillars – structural (*How can statistical profiling and rule-based logical checks be used to detect data quality issues and prepare water utility indicator data for reliable anomaly detection?*), modelling (*How can utility metadata and comparator-based modelling support the detection of deviations in performance indicators across diverse water utilities?*), prioritisation (*Which severity scoring methodologies can be investigated to best translate statistical deviations into a prioritisation of anomalies?*), technical validation (*To what extent does the proposed anomaly flagging system perform reliably, and align with expert validation and benchmarking expectations?*) and ethical review (*What ethical and institutional implications arise from implementing an automated anomaly detection framework in the context of global water utility benchmarking?*) – framing the development and evaluation of the system across all core dimensions.

4.2 Design & Methodological Framework

Rooted in principles of design science research, the approach balances technical modelling with ethical, contextual, and usability considerations.

Rather than adhering strictly to either agile[67] or waterfall methodologies[68], this thesis adopts a research-driven iterative prototyping cycle that bridges exploratory inquiry and technical implementation. This is rooted in the understanding that validation in global benchmarking is not purely a computational challenge, but a multidimensional problem shaped by data complexity, institutional variation, and normative expectations – all surfaced during the analysis in Chapters 2 and 3.

The following two subsections – Iterative Prototyping Cycle in Section 4.2.1, and Development Timeline in Section 4.2.2 – outline the backbone of this methodology.

4.2.1 Iterative Prototyping Cycle

To guide the structured yet adaptive development of the anomaly detection system, this thesis introduces a tailored Iterative Proto-

[67] **Agile Project Management:** An iterative approach that emphasises flexibility, stakeholder collaboration, and continuous improvement through short development cycles.

[68] **Waterfall Project Management:** A linear project management approach where each phase (e.g., planning, development, testing) is completed in sequence before the next begins.

typing Cycle displayed in Figure 4.1 – a custom-built methodology designed to balance technical rigour with human-centred responsiveness and ethical reflection.

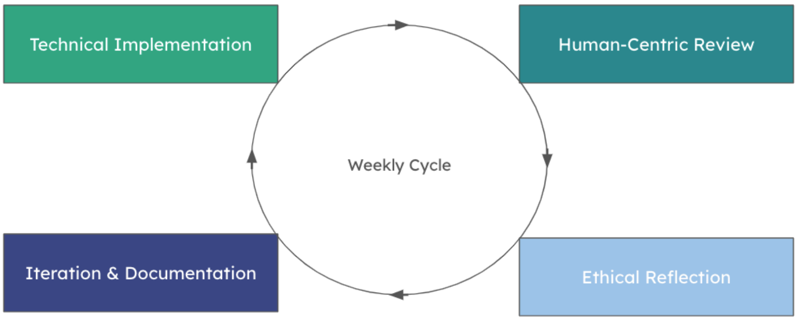


Figure 4.1: *The Iterative Prototyping Cycle used in this thesis, consisting of four interconnected stages – Technical Implementation, Human-Centric Review, Ethical Reflection, and Iteration & Documentation – guiding the continuous development and refinement of the anomaly detection framework.*

Unlike conventional engineering lifecycles, this approach is deliberately designed for the unique challenges of public-sector data systems, where institutional fragmentation, data uncertainty, and political sensitivity demand a more nuanced, reflexive design process. The cycle is conceptually grounded in van de Poel et al., 2011's *ethical cycle* and *design-for-values* principles, which highlight the iterative alignment between technological systems and normative goals; it also draws on Munro, 2021's *human-in-the-loop* paradigm, which emphasises the importance of continuous expert input in machine learning processes, and Tong et al., 2009's work on product innovation design, which advocates for iterative product development through stakeholder engagement and coordination across technical and functional domains.

Each iteration consists of four sequential but interlinked phases, repeated across core development sprints:

Step 1: Technical Implementation

This first phase centres on the systematic development of core system components, encompassing data preprocessing, algorithm design, and model configuration. Drawing on the technical insights reviewed in Chapter 3, each implementation decision reflects both prior research and contextual adaptation to the realities of benchmarking environments. Rather than building toward a static endproduct, this step prioritises exploratory experimentation and iterative refinement, allowing alternative modelling strategies to be tested, compared, and revised.

Step 2: Human-Centric Review

Once a technical implementation of the module is investigated, its outputs undergo human-centred review. This includes personal critical reflection on edge cases or unexpected results, as well as external feedback from domain experts – such as the World Bank's NewIBNET team and academic supervisors from Delft University of Technology. Their input provides essential contextual grounding, helping to assess not just whether the model works, but whether it aligns with institutional workflows.

Step 3: Ethical Reflection

Each cycle explicitly incorporates a reflective step focused on ethical implications. This goes beyond technical functionality to assess broader questions of fairness, transparency, and potential unintended consequences. The reflection considers both the current module and its contribution to the overall findings, forming a holistic ethical perspective that evolves alongside the technical design.

Step 4: Iteration and Documentation

Informed by insights from the previous stages, targeted refinements are made to the model logic or underlying assumptions. These changes are clearly documented, including the rationale behind them, the expert and ethical feedback that informed them, and how they position the system for the next development stage. Each cycle concludes with the formulation of an objective for the next iteration, ensuring focused, continuous progression.

> **Summary:** The iterative prototyping cycle offers a structured yet flexible approach, enabling the systematic development, testing, and refinement of anomaly detection strategies through 4 alternating phases: **Technical Implementation**, **Human-Centric Review**, **Ethical Reflection**, and **Iteration & Documentation**. Each iteration contributes to deeper insight into the interaction between automation, domain-specific constraints, and public-sector accountability.

4.2.2 Development Timeline

The development of this work followed a structured timeline spanning 7 months, from February 2025 to August 2025[69]. The pro-

[69] **Thesis Timeline:** For an in-depth overview of the timeline, see Appendix D.1.

cess was divided into three core phases: (1) **Literature Study & Gap Analysis**, (2) **Model Development & Validation**, and (3) **Synthesis & Recommendations**. Each phase builds incrementally on the preceding one: theoretical insights from the literature informed targeted system requirements; these, in turn, shape the design and evaluation of the automated framework; and finally, the outcomes of implementation are translated into analytical reflections and domain-relevant recommendations.

4.3 Critical Assumptions & Risk Analysis

The focus of this thesis is formed by both methodological constraints and institutional realities. While the work intends to introduce structural improvements to indicator validation processes, it has been designed to demonstrate feasibility within a limited timeline. Several design trade-offs and assumptions are made to balance rigour, usability, and scope, all of which carry implications for system performance and generalisability.

4.3.1 Critical Assumptions

The following assumptions underpin key design choices. Their invalidation could compromise the relevance or functionality of the system:

- **Sufficient Data Quality:** Despite gaps, it is assumed that submitted utility data remains representative enough to support anomaly modelling (Chandola et al., 2009).
- **Stable Indicator Definitions:** Indicator structures and units are assumed to remain consistent throughout the development period.
- **Stakeholder Input:** Timely feedback from the NewIBNET team and academic supervisors is assumed for iterative validation and contextual calibration.

4.3.2 Risk Analysis

Several risks were identified during the design phase, each influencing the experimental boundaries of this thesis:

- **Over-sensitivity to Sparse Data:** High rates of missing or anomalous inputs may distort flagging logic. To mitigate this, the system could look into incorporating gap detection, imputation safeguards, and controlled outlier thresholds (Clifton et al., 2022; Wu et al., 2021).

- **Static Comparator Limitations:** Fixed comparator groups offer interpretability but risk obsolescence in changing sociopolitical contexts. Dynamic benchmarking is excluded due to infrastructure constraints but remains a key area for future extension.
- **Misalignment with Human Review Practices:** Automated flags may not always align with reviewer judgment or operational relevance. This risk is partially mitigated through expert feedback loops and design of interpretable scoring logic (Baxter et al., 2011).
- **Ethical and Institutional Constraints:** Without downstream integration into decision processes, there is a risk that flags could reinforce existing inequities or be misapplied. While interpretive outputs remain out of scope, fairness principles are embedded through stratified comparators and transparency in scoring logic (Waller et al., 2025; McNealy, 2021).
- **Dataset Scope Limitation:** The experimental evaluation is based on a single dataset from the NewIBNET system, which may introduce contextual bias or limit the generalisability of findings. To mitigate this, supplementary testing is conducted using data from another utility association to assess transferability and robustness of the proposed approach.

4.3.3 Out-of-Scope Areas & Deferred Innovations

Certain capabilities – such as dynamic comparator calibration, test input separation protocols, and interactive reviewer tools – are deprioritised due to time and system complexity. Their exclusion does not signal irrelevance but reflects the focus on back-end robustness and institutional feasibility. Supplementary guidance materials[70] were developed in lieu of interface-level enhancements.

> **Summary:** Key assumptions included stable indicator definitions and stakeholder availability, while risks such as data quality issues and comparator misalignment were identified and addressed through mitigation strategies.

[70] Examples include reviewer documentation and a walk-through.

5 Data Preparation & Structural Validation

> **RQ1:** *How can statistical profiling and rule-based logical checks be used to detect data quality issues and prepare water utility indicator data for reliable anomaly detection?*

This chapter establishes the foundational integrity of the NewIBNET dataset through three sequential components: initial data exploration in Section 5.1, logical validation in Section 5.2, and missing data treatment in Section 5.3. It begins by assessing structural completeness, statistical variability, and reporting behaviours across raw inputs to determine whether the dataset supports robust downstream analysis. Next, it applies indicator-level validation rules to flag logically inconsistent entries. Finally, it addresses persistent data gaps in optional wastewater indicators through a comparative evaluation of imputation methods.

5.1 Initial Data Exploration & Characteristics

Before designing a scalable anomaly detection system, it is essential to first assess the structural integrity and statistical foundation of the raw dataset. This section outlines the initial data ingestion pipeline in Section 5.1.1 and evaluates the completeness, distributional properties, and reporting behaviour in Section 5.1.2. These steps serve two critical purposes: first, to ensure that the dataset provides a sufficiently robust statistical basis for downstream modelling, and second, to surface early-stage issues – such as missing values or implausible entries – that may justify immediate flagging. Such pipeline design and exploratory analysis establish the minimum data requirements for reliable validation and lay the groundwork for the rule-based checks and modelling logic that follow.

5.1.1 Pipeline Development

Given the absence of an existing technical stack, a local processing environment is set up using Python[71] – chosen for its flexibility, compatibility with tools like Power BI[72], and its suitability for scalable data manipulation. Confidential Excel files containing the full 2022, 2023, and partial 2024 utility submissions, totalling 289 utilities, are securely shared by the NewIBNET team.

As a first step, the raw inputs are imported into a code-friendly environment[73], where key preprocessing tasks can be performed: variable renaming for clarity, file reformatting for readability, and scripting basic data ingestion logic to transition away from opaque spreadsheet review. Initial inspection also revealed unit discrepancies – some fields used different measurement systems within the same column. These inconsistencies are addressed through a custom unit conversion mechanism[74], enabling alignment across utility submissions. In Figure 5.1, it can be seen exactly which raw data points required conversion.

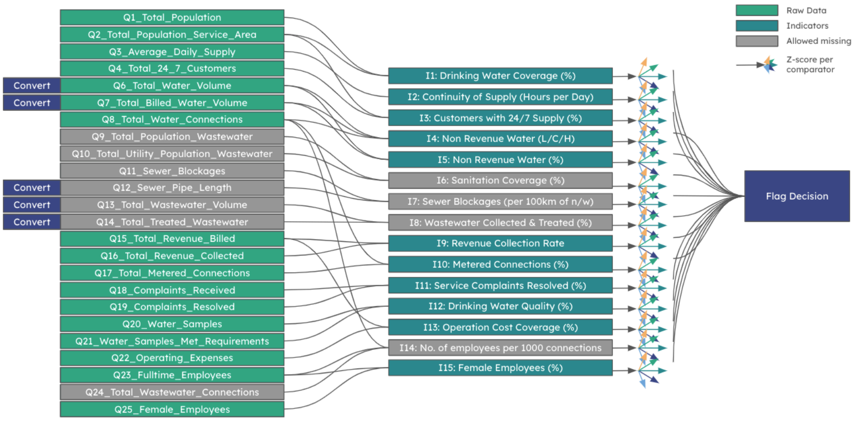

Q1_Total_Population
Q2_Total_Population_Service_Area
Q3_Average_Daily_Supply
Q4_Total_24_7_Customers
Convert
Q6_Total_Water_Volume
Convert
Q7_Total_Billed_Water_Volume
Q8_Total_Water_Connections
Q9_Total_Population_Wastewater
Q10_Total_Utility_Population_Wastewater
Q11_Sewer_Blockages
Convert
Q12_Sewer_Pipe_Length
Convert
Q13_Total_Wastewater_Volume
Convert
Q14_Total_Treated_Wastewater
Q15_Total_Revenue_Billed
Q16_Total_Revenue_Collected
Q17_Total_Metered_Connections
Q18_Complaints_Received
Q19_Complaints_Resolved
Q20_Water_Samples
Q21_Water_Samples_Met_Requirements
Q22_Operating_Expenses
Q23_Fulltime_Employees
Q24_Total_Wastewater_Connections
Q25_Female_Employees
I1: Drinking Water Coverage (%)
I2: Continuity of Supply (Hours per Day)
I3: Customers with 24/7 Supply (%)
I4: Non Revenue Water (L/C/H)
I5: Non Revenue Water (%)
I6: Sanitation Coverage (%)
I7: Sewer Blockages (per 100km of n/w)
I8: Wastewater Collected & Treated (%)
I9: Revenue Collection Rate
I10: Metered Connections (%)
I11: Service Complaints Resolved (%)
I12: Drinking Water Quality (%)
I13: Operation Cost Coverage (%)
I14: No. of employees per 1000 connections
I15: Female Employees (%)
Flag Decision
Raw Data
Indicators
Allowed missing
Z-score per comparator

Figure 5.1: *This illustrates the transformation pathway from raw numerical inputs (left) through derived indicators (centre), and into the comparator logic that determines final flagging outcomes (right). Raw data requiring unit conversion is marked with a dark blue "convert" square adjacent to the corresponding value.*

Each of the 289 utilities is expected to submit responses for questions Q1–Q25 (indicated in light green), with the exception of the wastewater-related inputs (indicated in grey). This is expected to result in a total of 6,936 data points – 24 per utility. These raw inputs are then used to compute 15 key indicators I1-I15 (indicated

[71] **Python:** A popular, high-level programming language known for its readability and versatility. See more: `https://www.python.org/`

[72] **Power BI:** A Microsoft data visualisation tool used to analyse and share interactive business insights – currently used in **Data Visualisation** layer of NewIBNET. See more: `https://www.microsoft.com/en-us/power-platform/products/power-bi`

[73] **Visual Studio Code:** A lightweight, open-source code editor developed by Microsoft, popular for programming across multiple languages. See more: `https://code.visualstudio.com/`

[74] This is needed for standardising water volumes and pipe length inputs.

in dark blue), yielding 4,335 derived indicator values across the dataset. These indicator values form the basis for comparison against relevant comparator groups, as will be discussed in the following chapters, to ultimately determine whether a data point is anomalous.

This preparatory phase sets the groundwork for deeper exploratory data analysis, which is essential for understanding the statistical behaviour and structural gaps within the dataset – going beyond surface-level observations to inform model design and flagging logic.

5.1.2 Exploratory Data Analysis

The second step involves conducting a statistical and structural assessment of the raw dataset, as part of the broader data preparation and validation pipeline. This phase aims to quantify the variability and sparsity of utility submissions across all raw inputs within the environment introduced in Section 5.1.1.

The rationale for this exploratory phase is well-supported in the literature, which emphasises the importance of statistical profiling as a means of understanding what constitutes *normal* behaviour before attempting to flag deviations (Chandola et al., 2009; Ahmed et al., 2016; Bhuyan et al., 2013)[75].

Descriptive Profiling of Raw Inputs

Each raw input field (Q1–Q25) is subjected to a standard statistical summary, including mean, median, minimum, maximum, and standard deviation. These metrics are computed across all available utility entries to assess dispersion and identify potential anomalies in scale or input range. For instance, in the case of population service size (Q2), a wide spread with 1 as the minimum, ~190k as the median, and a maximum of 22 million, spanning several orders of magnitude, is observed.

In addition to this, a type-check verifies that each raw input matches the expected numerical format. Since some test data is reportedly submitted outside the standard survey system, the code also attempts to convert string-formatted numbers, flagging any entries failing to convert. No type errors are found among the 289 utilities.

Missing Data and Placeholder Analysis

Alongside descriptive metrics, a gap analysis is conducted to determine the completeness of each raw variable. This involves cal-

[75] As Chandola et al., 2009 stated, *"...anomalies are patterns in data that do not conform to a well-defined notion of normal behavior."* – a concept that requires baseline characterisation of each input variable.

culating the total number of null entries per field and examining patterns of missingness across utilities. A particularly salient issue is the appearance of the placeholder value '1.0', which is identified – both by NewIBNET reviewers and confirmed through manual inspection – as a non-informative default used by some utilities during data entry. Since '1.0' is not a plausible value for any of the raw inputs, it is reclassified as a missing value. To ensure consistency, the descriptive metrics presented earlier are recalculated after this adjustment so that distributions reflect only valid entries. Out of 6,936 expected entries, a total of 975 are missing, and 248 contain a placeholder value of '1.0', resulting in approximately 17% of the dataset being incomplete.

The results, presented fully in the Appendix A.2, further reveal significant sparsity in wastewater-related indicators, with missing values ranging from 131 to 139 for each raw input Q9, Q10, Q11, Q12, Q13, Q14, and Q24 – all classified as optional wastewater-related survey questions. However, inconsistencies are also found in mandatory non-wastewater fields, such as total water connections (Q8) and full-time employees (Q23), suggesting data entry issues beyond optional service non-provision.

Distribution Visualisation

Understanding the distribution of input data is a crucial first step in anomaly detection, as it provides insight into the underlying patterns, variability, and potential irregularities in the data. Examining whether values are normally distributed, skewed, or multimodal helps to differentiate expected variation from deviations that require further scrutiny. Histogram-based visualisation and parametric statistical modelling are commonly employed for this task (Chandola et al., 2009; Wu et al., 2021), and their implementation here form the empirical foundation for subsequent rule-based checks and comparator-based modelling.

Univariate histograms are generated for key variables, starting with population service size (Q2) and total water connections (Q8). As shown in Figure 5.2, both distributions exhibit strong right skewness, characterised by a high concentration of small-scale utilities and a long tail of larger urban providers.

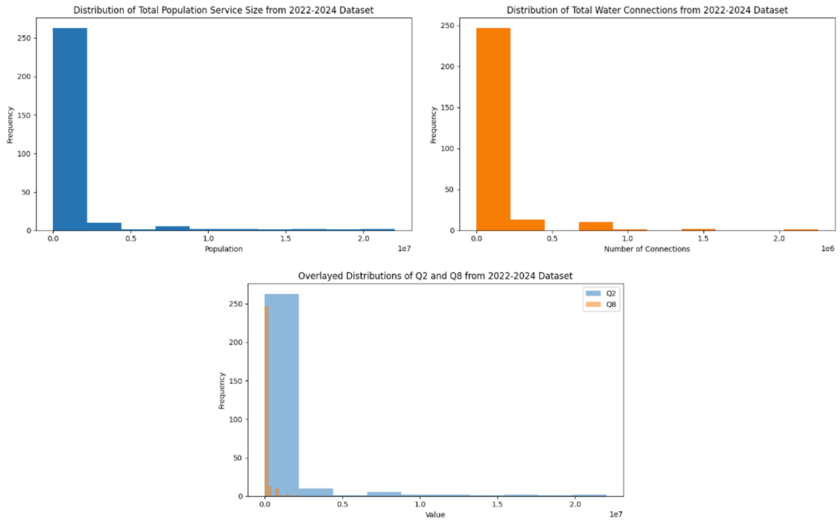


Figure 5.2: *All three histograms illustrate the distributions of population service size (Q2), total water connections (Q8), and a combined overlay of both – where Q2 is shown in blue and Q8 in orange. The x-axis represents the number of people served (up to a magnitude 10 million) and the number of water connections (up to magnitude 1 million), reflecting the wide variability and exponential scale of utility sizes. The y-axis indicates the frequency of utilities within each population or connection range.*

This exponential pattern indicates the need for stratification, as applying a single global threshold could risk obscuring legitimate local variations and disproportionately penalising smaller utilities.

This insight is reinforced in Figure 5.3, which plots the population service size (Q2) across predefined World Bank population categories. Within each segment, distributions appear more uniform and better suited to z-score based profiling. These findings support the segmentation approach proposed in the work of Manghee et al., 2012 and Bhatt, 2024, and later operationalised in the comparator logic described in Chapter 6.

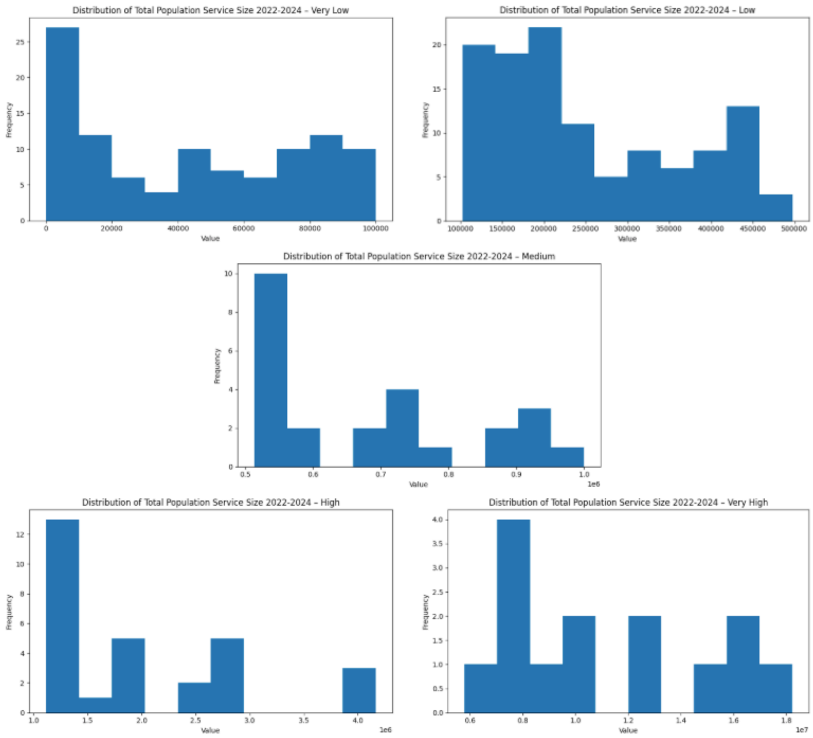


Figure 5.3: *This figure presents five separate histograms, each showing the distribution of population service size within a specific World Bank-defined category: Very Low (0–100,000), Low (100,001–500,000), Medium (500,001–1,000,000), High (1,000,001–5,000,000), and Very High (5,000,001 and above). By segmenting the data into these predefined ranges by the World Bank, the figure allows for a clearer comparison of the internal variation within each category and highlights how population sizes are distributed across different utility scales.*

Identification of Structural Artefacts

Finally, several implausible values are flagged during manual review. The most extreme case involves a reported total population size (Q1) exceeding 2 billion – an artefact likely stemming from internal spreadsheet testing. The existence of such entries confirms a lack of internal filtering mechanisms and indicates the need for logical rule-based validation, which is developed in the following section. These findings echo earlier research on metadata validation and structural constraints (Raciti et al., 2012; Y. Liu et al., 2025).

In total, this process identified 21 utilities with missing data in fields where values are expected. These 21 utilities are now excluded from further analysis, as the incomplete records suggest either test data or submissions not suitable for deeper validation.

> **Summary:** This phase established a foundational data pipeline and conducted a structural and statistical assessment of the raw dataset, revealing key issues such as unit inconsistencies, missing values, and structural outliers.

5.2 Logical Validation Rules

Logical validation rules assess the internal coherence of the 15 derived indicator values – computed metrics that form the foundation of NewIBNET's benchmarking process. While the initial data exploration addresses surface-level issues such as missing values, placeholder entries, and data type mismatches, this stage moves deeper by embedding logical and mathematical checks into the indicator calculations themselves. These checks are grounded in what we *know* and *expect* from well-formed utility data – flagging implausible ratios, invalid percentages, and undefined operations such as division by zero.

In doing so, logical validation shifts the focus from basic data hygiene toward verifying whether reported values adhere to the fundamental principles behind each indicator.

5.2.1 Validation Logic & Design

Each of the 15 core indicators is computed using fixed combinations[76] of raw inputs, with validation rules applied during the computation process. Drawing from sector-specific norms and statistical literature (Wu et al., 2021; Abián et al., 2019), indicators are evaluated for:

- **Percentage bounds:** Most percentage-based indicators are logically constrained to a 0–100% range. Exceptions, such as wastewater treatment coverage (I8) and operational cost coverage (I13), are permitted to exceed this due to inflows or financial surpluses, respectively.
- **Division-by-zero protection:** Any operation involving division includes a pre-check to ensure denominator validity.
- **Non-negativity constraints:** Indicators relying on subtractive logic[77] are checked to ensure outputs remain within realistic bounds.
- **Input completeness:** Indicators depending on missing or flagged raw values are withheld from final computation and routed

[76] **Indicator Equations:** Refer to the overview provided in Appendix A.1

[77] **Subtractive Logic:** Non-revenue water indicators I4 and I5 use subtractive logic.

for review.

5.2.2 Utility Dataset Results

Applying these validation rules to the dataset reveals a range of indicator-level anomalies. In total, 36 instances among 29 utilities fail to meet the validation criteria outlined in Section 5.2.1. A detailed breakdown of the results is presented in Table 5.1.

Table 5.1: *Overview of Detected Logical Validation Errors*

Type of Error	Number of Violations
Percentage bounds exceeded	27
Negative values in subtractive indicators	6
Division-by-zero	3
Total	**36**

Notably, 27 of these violations, accounting for 75% of logical issues, are due to percentage values exceeding 100% in indicators that are logically capped. These flags point to fundamental errors in data entry or indicator computation, which would not have been identified through surface-level structural checks alone.

Unusual patterns also emerged in indicators that are technically permitted to exceed 100%. For instance, values for operational cost coverage (I13) typically range between 100–150%, but several extreme outliers far exceed 1000%. While such figures may be plausible in exceptional cases – such as utilities involving post-subsidy revenues[78] – they raise interpretive concerns when no corresponding explanation for the surplus is provided. Similarly, wastewater collection ratios (I8) occasionally exceed 100%, which could reflect inflow or infiltration[79] effects, but in most instances warrant further clarification to ensure data credibility.

Looking beyond violations detected through validation rules, the indicator summary sheet[80] highlights extreme outliers that depart sharply from sector norms[81]. Most indicators – such as drinking water coverage (I1), continuity of supply (I2), and non-revenue water in percentage terms (I5) – generally fall within expected performance ranges for low- and middle-income contexts (World Bank, 2014). However, several metrics contain values that far exceed operational plausibility: non-revenue water (I4) exceeding 16 million[82], sewer blockages (I7) surpassing 480,000 per 100 km, collection rates (I9) reported at over 600,000%[83], and operational cost coverage (I13) above 58 billion%[84]. Such magnitudes are more con-

[78] **Post-subsidy Revenues:** Refers to external financial inflows that may be excluded in operating revenues, leading to ratios above 100% (Foster et al., 2010).

[79] **Inflow and Infiltration:** Refers to unintended stormwater (inflow) or groundwater (infiltration) entering the sewer system through faulty pipes, manholes, or illegal connections (EPA, 2014).

[80] Full indicator descriptive statistics overview can be found in Appendix A.4.

[81] Sector norms here refer to the IBNET-recorded standards from 2014 (World Bank, 2014).

[82] Typical IBNET Range: 50–60 m³/km/day (World Bank, 2014).

[83] Typical IBNET Range: median of 85–105% (World Bank, 2014).

[84] Typical IBNET Range: < 100% (World Bank, 2014).

sistent with unit misalignment, data entry errors, or unfiltered raw values than with genuine performance variation. Even when medians lie within credible bounds, these extreme cases indicate the need for automated plausibility limits and targeted follow-up to safeguard the integrity of benchmarking outputs.

Additional anomalies include negative values in subtractive indicators, accounting for 6 of the 36 total violations, as well as isolated division-by-zero errors, which occurs in 3 cases. Both types of issues are automatically flagged and excluded from further analysis.

A total of 29 utilities are flagged during this stage, in addition to the 21 utilities previously identified in the exploratory data analysis. This indicates the layered design of the pipeline, where structural and logical validations operated in tandem to catch both surface-level and embedded inconsistencies.

> **Summary:** This introduced a suite of logical validation rules designed to assess the mathematical coherence and conceptual plausibility of derived indicator values. By flagging violations such as percentage overflows, negative ratios, and division-by-zero errors, this stage ensured that data integrity extends beyond surface-level completeness to include internal consistency, reinforcing the reliability of subsequent anomaly detection and benchmarking.

5.3 Missing Data Treatment: Imputation

Persistent data gaps, particularly in wastewater indicators, pose a fundamental challenge to the integrity and completeness of anomaly detection. Rather than excluding partially submitted entries, this section explores whether statistical imputation can be used to recover plausible values and preserve analytical coverage.

Drawing on prior literature, two methods are evaluated: a parametric statistical approach using median imputation (Miao et al., 2024), and a k-nearest neighbour (k-NN) approach adapted from privacy-preserving data systems (Clifton et al., 2022). Each method is tested on a selection of utilities with wastewater reporting gaps to assess their impact on downstream anomaly detection results. Importantly, imputed values are treated as diagnostic intermediaries only – used internally to support fairer flagging coverage, but not to retain or publish in the final **Data Visualisation** stage.

5.3.1 Eligibility Criteria for Imputation

Imputation in this study is applied only to wastewater indicators as this subset shows inconsistent gaps[85], where some utilities report certain wastewater data but not others, creating ambiguity. Rather than assuming the absence of wastewater services, imputation can be used here to give utilities the benefit of the doubt, acknowledging that a utility may provide wastewater services but lack data for specific measures. The focus of this experiment is therefore to test whether retaining such utilities, with statistically grounded imputation applied for review purposes, strengthens validation. By contrast, imputing non-wastewater indicators is more problematic given their weaker interlinkages, and is therefore left out for now.

All utilities are classified based on the availability of raw data required to compute four key wastewater indicators: sanitation coverage (I6), sewer blockages (I7), wastewater collected and treated (I8), and number of employees per 1000 connections (I14). Each utility is assigned to one of four categories:

- **Completely Missing:** No data available for any of the four wastewater indicators. These are assumed to represent non-wastewater service utilities, and no imputation is applied.
- **Completely Filled:** All four indicators could be computed from available data. These cases require no intervention and are excluded from the imputation process.
- **Sporadic:** Only one of the four indicators is derivable, raising concerns about data quality or reporting inconsistencies. These cases are flagged for reviewer inspection but not imputed.
- **Eligible for Imputation:** Two or three indicators could be derived, indicating partial but coherent reporting. These submissions are considered suitable for imputation.

Table 5.2 shows the distribution of utilities across imputation classification groups within the 2022–2024 NewIBNET dataset.

[85] **Inconsistent Wastewater Gaps:** Raw data inputs Q9, Q10, Q11, Q12, Q13, Q14, and Q24 each had around 131–139 missing values. Further details are provided in Appendix A.2.

Table 5.2: *Imputation Classification Groups for the 2022-2024 NewIBNET Dataset*

Category	Number of Utilities
Completely Missing	127
Completely Filled	108
Sporadic	9
Eligible for Imputation	24
Total	268

Only the fourth group of **Eligible for Imputation**, comprising 24 utilities and 32 instances, is selected for experimental comparison using both median and k-NN approaches. The evaluation compares indicator behaviour in the **Eligible for Imputation** group against the **Completely Filled** baseline.

5.3.2 Median Approach

As a simple approach, median imputation provides a transparent, low-complexity method for estimating missing values. This technique replaces missing entries with the median of the corresponding indicator across comparator groups, offering a robust estimate that resists distortion from outliers (Miao et al., 2024). The median values for each wastewater indicator are presented in Table 5.3.

Table 5.3: *Overview of Median Imputation Results for Wastewater Indicators within the NewIBNET 2022-2024 Dataset*

Indicator	Number of Imputed Values	Median Value
I6	1	54.000
I7	10	225.616
I8	17	100.000
I14	4	4.597

5.3.3 k-NN Approach

Inspired by the methodology proposed in Clifton et al., 2022, this experiment applies a k-NN approach to impute missing wastewater indicators by identifying utilities with similar profiles across known variables. This method is particularly suited to NewIBNET's structure, where utility characteristics vary substantially but appear to fit more evenly into predefined population clusterings. This method results in different imputed values per instance, unlike median imputation which produces a single value for all indicators.

To first assess the appropriate number of neighbours (k) for k-NN imputation, a comparative sensitivity analysis is conducted across values k = 2 to 7[86]. For each setting, the imputed distributions are overlaid with the distribution of utilities with complete (non-imputed) indicator values to evaluate alignment in shape, central tendency, and spread. The objective is to identify which value of k best preserves the expected statistical behaviour derived from neighbouring utilities, thereby yielding the most contextually faithful imputations. The results can be seen in Figure 5.4.

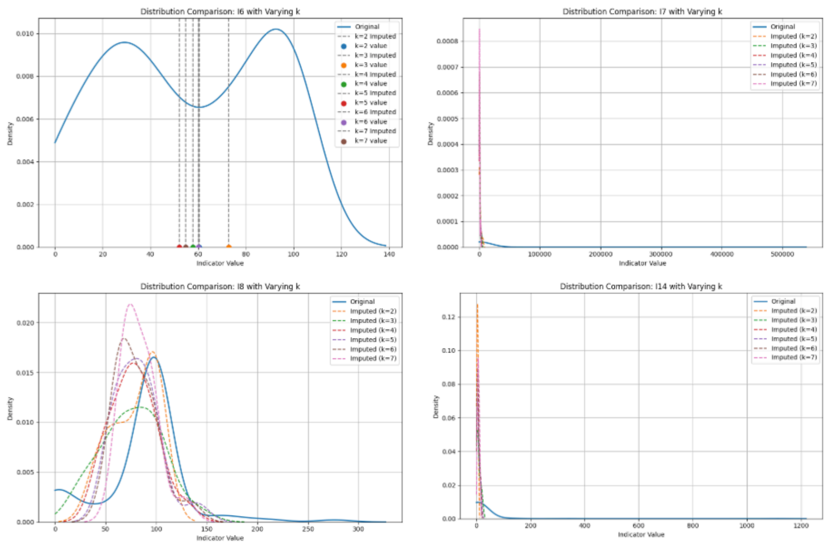


Figure 5.4: *This figure presents the sensitivity analysis for k-NN imputation, evaluating values of k ranging from 2 to 7. It consists of four graphs, each corresponding to one of the wastewater indicators under examination: I6 (top left), I7 (top right), I8 (bottom left), and I14 (bottom right).*

Overall, no major divergence is observed in sewer blockages (I7) and number of employees per 1000 connections (I14), with distribution curves showing substantial overlap across all tested values of k. This indicates relative insensitivity to parameter variation. In contrast, wastewater collected and treated (I8) demonstrates clearer variation, with k = 2 and k = 4 exhibiting slightly better alignment with the original distribution than the default k = 3. For sanitation coverage (I6), using k = 3 results in imputed values that aligned more closely with the second peak of the bimodal distribution, indicating a stronger contextual fit with high-coverage utilities compared to higher k-values, which averaged across modes

[86] Small k values (≈2–10) are often tested in k-NN imputation for datasets with limited structure, e.g. Troyanskaya et al., 2001 on microarray data.

and shifted the result into the low-density valley.

k = 3 is retained for consistency in subsequent analysis, given the absence of strong performance differentials in most indicators. Nevertheless, this experiment shows an important insight: the optimal value of k may be indicator-specific rather than global. Tailoring k per indicator, potentially informed by cross-validation or structural similarity metrics, could further improve the contextual accuracy of imputation within a given dataset and warrants future exploration.

5.3.4 Performance and Robustness Comparison

This section compares the outputs of median and k-NN imputation across the four wastewater indicators. Each imputed distribution is evaluated against the distribution of utilities with complete (non-imputed) indicator values as shown in Figure 5.5.

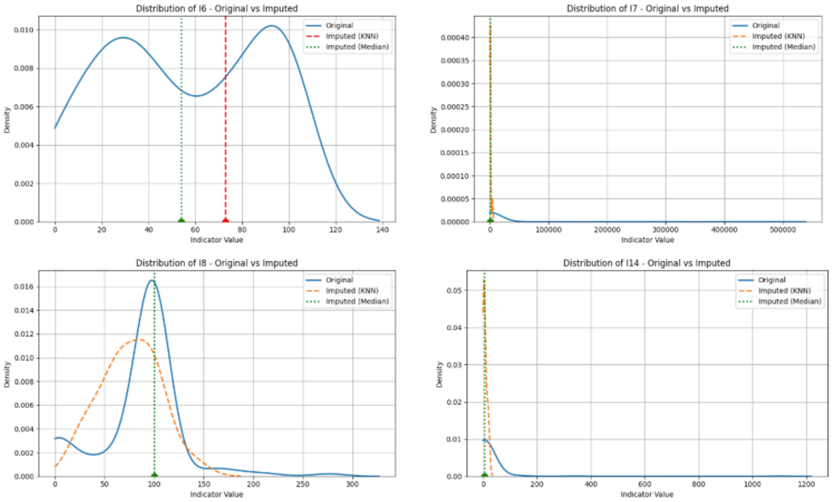


Figure 5.5: *Distribution of wastewater indicators comparing original baseline distributions (based on complete entries, shown in blue), k-NN imputation results with k = 3 (shown in red and orange), and median imputation results (shown in green).*

Sanitation coverage (I6) presents a distinct bimodal distribution over a 0–140% range, with peaks around 30% and 90%, reflecting the divide between utilities with limited versus extensive infrastructure. The k-NN-imputed value (~70%) is positioned centrally between these peaks – a region of low empirical density, suggesting the method blends neighbours from both clusters. Median imputation (~55%) also falls into this dip, indicating the challenge both methods face when reconciling multimodal patterns. While k-NN benefits from contextual grounding, its performance is difficult to assess for this indicator, as only a single value required imputation; neither approach fully captured the real-world segmentation, highlighting the need for reviewer oversight in heterogeneous distributions.

In contrast, wastewater treated (I8) follows a unimodal, right-skewed distribution, concentrated between 70–100 units. Here, k-NN closely replicated the original curve, peaking slightly earlier (~70) and tapering off conservatively before 180. Median imputation (at 100) coincides with the empirical peak, making it statistically optimal in centrality but less reflective of the natural spread. k-NN provided a better match to shape and dispersion, reinforcing its strength in continuous, moderately skewed settings.

Both sewer blockages (I7) and employee ratios (I14) exhibits highly right-skewed distributions, with dense clustering at low values (0–40) and sparse long tails reaching up to 500,000 and 1,200, respectively. In both cases, k-NN restricts imputed values to the populated low-value region, avoiding tail extrapolation and preserving distribution realism. Median imputation yields values near the same peak, but without any variation. This highlights k-NN's strength in avoiding distortion while preserving nuance, even in sparse or extreme-value contexts.

From the perspective of anomaly detection, the choice of imputation method has direct implications for the reliability and interpretability of downstream statistical analyses. Median imputation, while robust against outliers, imposes a fixed central tendency that can obscure genuine variability, which can be particularly problematic when imputing data from utilities that may exhibit anomalous but contextually valid behaviour. In contrast, k-NN imputation adapts to local data structure, allowing imputed values to reflect the variability and contextual signature of their nearest neighbours. Although k-NN may reinforce erroneous patterns if neighbouring data points are themselves flawed, it does so in a probabilistically consistent manner, preserving the empirical relationships within the data.

Therefore, in the context of flagging where outliers may represent either errors or legitimate edge cases, k-NN offers a more nuanced foundation for anomaly flagging, especially when paired with diagnostic visualisations and expert oversight.

Summary: This section explored whether statistical imputation can enhance coverage of partially reported wastewater indicators without compromising anomaly detection, comparing median-based and k-nearest neighbour (k-NN) methods. While median imputation offered simplicity, k-NN better preserved contextual variability and distributional nuance, making it a more suitable foundation for reliable, nuanced anomaly flagging in incomplete datasets.

6
Context-Aware Anomaly Modelling Frameworks

> **RQ2:** *How can utility metadata and comparator-based modelling support the detection of deviations in performance indicators across diverse water utilities?*

To meaningfully assess whether a utility's indicator values reflect anomalous behaviour, they must be interpreted relative to a suitable context. This chapter investigates how comparator groups – predefined peer sets used to establish normative baselines – could influence the fairness and sensitivity of anomaly detection outcomes. Overview of Comparator Design in Section 6.1 outlines the logic behind comparator design; Existing Static Comparators in Section 6.2 analyses the legacy static comparator groups currently embedded in the NewIBNET framework; and New Static Comparators in Section 6.3 introduces two newly proposed comparator configurations, developed and examined in this thesis to improve contextual validity.

6.1 Overview of Comparator Logic

In the manual NewIBNET review process, a comparator group refers to the population over which a given comparator type is calculated, forming the internal reference point against which utilities' indicator values are assessed for potential anomalies. While the term *comparator group* is used here as an analytic construct, it aligns with the implicit logic already present in NewIBNET's manual reviews.

Each of the 15 indicators calculated per utility is compared to the values within its assigned comparator group. This is operationalised through z-score calculations:

$$z = \frac{x - \mu}{\sigma}$$

where x is the utility's indicator value (I1-I15), μ is the comparator group mean, and σ is the group's standard deviation. The resulting z-score provides a consistent way to measure how far a utility's value deviated from the average of its comparator group.

> **Summary:** This section outlined how static comparator groups were used in the NewIBNET review process to assess utility indicators by calculating z-scores, enabling standardised detection of deviations from group norms.

6.2 Existing Static Comparators

The existing manual NewIBNET validation framework relies on three primary static comparator types: the **global average**, **population service size**, and **the number of water service connections**. Each of these offer a structured basis for assessing utility performance relative to peers, serving as internal reference points during the **Data Review** process.

6.2.1 Global Average Comparator

This comparator involves calculating a single mean and standard deviation for each indicator across all utilities in the dataset. As introduced in Chapter 2, this serves as the most generic reference point and is likely selected for its simplicity and wide applicability. However, while appealing in its universality, Chapter 5 quickly reveals limitations in heterogenous datasets like NewIBNET. Utilities from vastly different operational contexts are treated as comparable, potentially resulting in distorted or overly broad interpretations of deviation.

6.2.2 Population Service Size Comparator

To improve contextual relevance, utilities are grouped according to the population service size (Q2), following predefined thresholds established by the World Bank. Specifically, each utility is assigned to one of five population-based categories[87]: *Very Low* (0–100,000), *Low* (100,001–500,000), *Medium* (500,001–1,000,000), *High* (1,000,001–5,000,000), and *Very High* (5,000,001 and above). This stratification enables performance comparisons among utilities with comparable demographic and operational scale, recognising that factors such as infrastructure capacity, service coverage, and resource constraints often vary systematically with population size. The distribution of utilities across these categories, based on the NewIBNET 2022–2024 submissions, is presented in Table 6.1.

Table 6.1: *Overview of the Population Service Size Comparator Groups of the NewIBNET 2022-2024 Dataset*

Population Category	Number of Utilities
Very Low	104
Low	115
Medium	25
High	29
Very High	16
Total	**289**

Within this distribution overview, it becomes evident that some categories are more densely populated than others. For example, the *Very Low* and *Low* categories contains the highest number of utilities, with 104 and 115 respectively, while the remaining categories are significantly smaller, ranging between 16 to 29 utilities. This aligns with the findings from Chapter 5 on statistical profiling, which revealed a strong right-skewed distribution. Most utilities are smaller in size and therefore fall into the lower categories of the defined range.

6.2.3 Water Service Connections Comparator

A parallel stratification approach is applied based on the total number of water service connections (Q8) reported by each utility. As with population size, utilities are assigned to one of five predefined categories established by the World Bank: *Very Low* (0–10,000), *Low* (10,001–50,000), *Medium* (50,001–100,000), *High* (100,001–200,000), and *Very High* (200,001 and above). This dimension serves as an al-

[87] A visual representation of the distribution of the population-based categories is provided in Chapter 5, Figure 5.3.

ternative proxy for size, reflecting the infrastructure footprint and connection density of each utility. The distribution of utilities across these service connection categories, based on the NewIBNET 2022–2024 submissions, is shown in Table 6.2.

Table 6.2: *Overview of the Water Service Connections Comparator Groups of the NewIBNET 2022-2024 Dataset*

Connections Category	**Number of Utilities**
Very Low	65
Low	132
Medium	39
High	11
Very High	28
Total	**275**

The same trends observed in the population comparator are evident here as well. The *Very Low* and *Low* categories comprised a much larger group compared to the *High* and *Very High* categories, once again reflecting the right-skewed nature of the current data.

> **Summary:** This section described the three static comparators used in the NewIBNET validation framework – global average, population service size, and number of water connections. The distribution of utilities across these categories revealed a strong right skew, with most falling into the *Very Low* and *Low* groups, highlighting the importance of using stratified comparators for fairer anomaly detection.

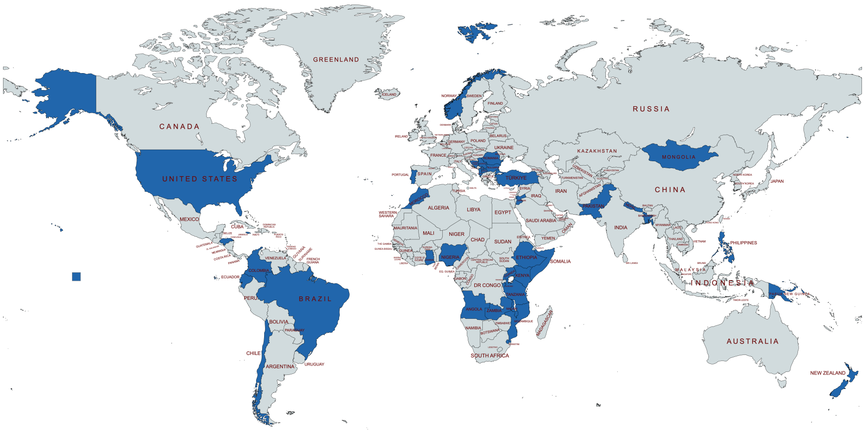
6.3 New Static Comparators

While the existing static comparators provide a foundational mechanism for contextual benchmarking, this thesis also explores whether alternative groupings could enhance the fairness and sensitivity of deviation analysis. In particular, the classification systems developed by the World Bank, 2024, including **regional** and **income-based** groupings[88], offer a natural extension. These taxonomies are widely recognised in global development discourse and are already integrated into NewIBNET's institutional framework, making them both relevant and operationally compatible. Their inclusion also responds to insights from the literature in Section 3.2.7,

[88] The **World Bank Country and Lending Groups** classification is updated annually on July 1 (World Bank, 2024). This thesis applies the 2024–2025 fiscal year classification.

where IBNET is frequently employed as a foundation for comparative policy and performance research – from evidence-based political economy analysis (Manghee et al., 2012) and African utility assessments (C. v. d. Berg et al., 2017) to tariff subsidy evaluation (Andrés et al., 2020), fraud detection frameworks (Detroz et al., 2017), and financial-service quality studies (Tsagarakis, 2018). At the same time, critiques such as Bhatt, 2024's work remind us that efficiency-centric benchmarks risk obscuring local equity concerns, highlighting the need for more nuanced comparator segmentation.

A visual map of participating utilities in Figure 6.1 illustrates the broad geographic distribution – spanning multiple continents – that underpins the need for differentiated analysis.



Created with mapchart.net

Figure 6.1: *This map highlights the countries that participated in the 2022–2024 data intake, with all blue-coloured countries actively contributing to the database.*

Figure 6.2 presents Utility X as an example, illustrating how a single utility is assessed for comparative analysis on one indicator (I1), and how five comparator results can be used to identify deviations within their respective peer groups.

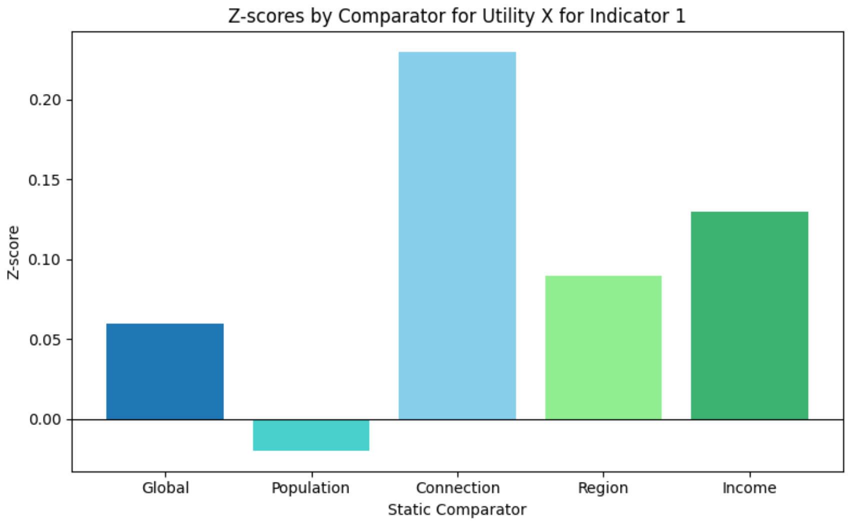


Figure 6.2: *This figure displays Utility X's z-scores on Indicator I1 across different comparator groups. The x-axis shows the existing and new static comparators – global average, population service size, number of water connections, World Bank Region group, and World Bank Income group – while the y-axis shows the corresponding z-scores for Utility X, calculated relative to each comparator distribution. The utility's name has been anonymised for privacy reasons and is referred to here as Utility X.*

6.3.1 World Bank Region Comparator

This comparator stratifies utilities according to the seven official World Bank regions, based on the World Bank's annual country and lending group classifications (World Bank, 2024). Each utility is mapped to a region according to the country in which it operates. The predefined regional categories are:

1. East Asia and Pacific,
2. Europe and Central Asia,
3. Latin America and the Caribbean,
4. Middle East and North Africa,
5. North America,
6. South Asia, and
7. Sub-Saharan Africa.

In theory, this allows the system to flag anomalies within region-specific expectations, rather than applying a one-size-fits-all standard globally.

However, initial findings from the 2022–2024 dataset highlight a similar yet critical pattern: distributional imbalance. The exact distribution of utilities for this comparator, based on the NewIBNET 2022–2024 datasets, is shown in Table 6.3.

Table 6.3: *Overview of the World Bank Region Comparator Groups of the NewIBNET 2022-2024 Dataset*

Region	Number of Utilities
East Asia and Pacific	25
Europe and Central Asia	40
Latin America and the Caribbean	14
Middle East and North Africa	7
North America	1
South Asia	37
Sub-Saharan Africa	161
Total	**285**

Nearly 57% of all utilities in this submission round comes from *Sub-Saharan Africa*, followed by 14% from *Europe and Central Asia*. In contrast, *North America* and the *Middle East and North Africa* are significantly under-represented. These imbalances raise important questions about the statistical robustness of regional comparators and their applicability when sample sizes are low.

6.3.2 World Bank Income Comparator

This comparator classifies utilities into one of four income categories as defined by the World Bank:

1. Low Income,
2. Lower-Middle Income,
3. Upper-Middle Income, and
4. High Income.

Income groupings offers a complementary lens to regional comparators, capturing economic factors that may influence investment capacity, tariff structures, labour markets, and service quality.

The exact distribution of utilities for this comparator, based on the NewIBNET 2022–2024 datasets, is shown in Table 6.4.

Table 6.4: *Overview of the World Bank Income Comparator Groups of the NewIBNET 2022-2024 Dataset*

Income Level	Number of Utilities
Low	88
Lower-Middle	137
Upper-Middle	40
High	21
Total	**286**

As with the regional comparator, utility distribution is uneven. The majority of utilities in the 2022–2024 dataset fall into the *Low* and *Lower-Middle* Income categories, with less than 10% coming from *High* Income countries. While this highlights NewIBNET's mission to serve emerging and developing contexts, it also poses statistical limitations when applying these comparators to underrepresented groups.

The introduction of additional comparator dimensions aims to broaden the contextual basis for anomaly interpretation, offering a way to explore whether structural and geopolitical variation – such as region or income level – can supplement existing groupings by population and connection size.

However, this expansion also raises new challenges. The uneven distribution of utilities across comparator groups highlights that more context does not necessarily yield more reliable insight. Comparisons drawn from sparsely populated groups or those with high internal diversity may introduce as much noise as clarity. This brings fourth a critical consideration: if comparator groups vary in representativeness and reliability, *should their diagnostic weight be adjusted accordingly?*

> **Summary:** New Static Comparators explored the use of alternative static comparators – based on World Bank regional and income classifications – to improve the contextual relevance of anomaly detection. While these dimensions offered valuable insight into structural and geopolitical diversity, their uneven group distributions highlighted limitations in statistical robustness.

7 Severity Scoring & Decision Framework

> **RQ3:** *Which severity scoring methodologies can be investigated to best translate statistical deviations into a prioritisation of anomalies?*

Having established the structure and rationale behind comparator groupings in Chapter 6, the next challenge is determining how deviations from these groups should be interpreted and acted upon. The preceding chapter closed with a critical question: if comparator groups differ in statistical robustness and representativeness, *should their diagnostic influence be weighted accordingly*?

Existing Threshold Logic in Section 7.1 first reviews the legacy framework's static z-score thresholds and flagging logic. Weighting in Section 7.2 then introduces a revised model that weighs comparator scores based on internal variance and group size. Finally, Severity Flagging in Section 7.3 proposes an integrated decision framework that combines comparator weights, score aggregation, and severity tiers to better align flagging outcomes with statistical and contextual credibility.

7.1 Existing Threshold Logic

Before proposing a refined scoring model, it is important to examine the threshold-based logic within the current NewIBNET system. Once all 15 indicator values are calculated, each is compared against three predefined static comparators: global average, population service size, and number of water service connections. A utility is flagged if the z-score for any indicator falls outside the range $[-4, 4]$. This threshold intends to capture extreme deviations without triggering flags for contextually acceptable variation.

While this approach provides a simple and transparent starting point, it embeds a critical limitation: each comparator is treated with equal diagnostic weight, irrespective of its statistical robustness. Yet, as Chapter 6 highlighted, the distribution of utilities across comparator groups is highly uneven, with some categories containing far fewer utilities than others. These imbalances directly signal the need to incorporate comparator reliability into the anomaly detection process, rather than assuming uniform diagnostic value.

> **Summary:** This section examined the current threshold-based system, which flags indicator values as anomalies if any z-score across three equally weighted static comparators falls outside the range $[4, 4]$, highlighting its simplicity but also its limitation in overlooking differences in statistical reliability across comparator groups.

7.2 Weighting

Building on the limitations identified in the existing threshold logic, this section explores how weighted scoring could enhance the fairness and statistical rigour of anomaly detection. The central premise is that not all comparator groups offer equal diagnostic reliability: some exhibit tighter, more stable distributions, while others are sparse or highly variable, thereby weakening confidence in their reference statistics. To address this imbalance, the intention is to assign weights to each comparator and multiply these with their corresponding z-scores, ensuring that individual deviations are scaled in proportion to the statistical reliability of the group from which they are derived. Drawing on principles from decision science and anomaly scoring literature (Kim et al., 2024; Hajirahimi et al., 2023), this phase experiments with multiple severity weighting schemes.

Two similar weighting approaches are developed and tested in this thesis, incorporating the newly introduced World Bank Region and Income comparators into the validation process. These models aim to reduce false positives, better reflect contextual reliability, and offer a more principled basis for evaluating utility performance.

7.2.1 Simple Weights

As a first approximation, comparator weights are derived based on the relative variance of z-scores across all indicators. For each comparator i, a global weight is computed as:

$$\text{Weight}_i = 1 - \frac{\text{Var}(Z_i)}{\sum_{j=1}^{5} \text{Var}(Z_j)}$$

Where $\text{Var}(Z_i)$ denotes the variance of z-scores produced by comparator i across the full dataset. The denominator is the total variance across all comparators.

This provides a coarse measure of comparator consistency, with lower variance suggesting greater statistical reliability. Five weights – one for each comparator – are calculated with the intention of applying them uniformly across all 15 indicator z-scores for each utility. The results can be seen in Table 7.1.

Table 7.1: *Simple Weight Results for the NewIBNET 2022-2024 Dataset*

Comparator	Weight
Global	0.7894
Population	0.8026
Connections	0.7977
Region	0.8082
Income	0.8021

Under this model, the World Bank Region comparator emerges as the most stable with 0.8032, while the Global Average exhibits the highest variability with weight 0.7894.

However, this approach has key limitations. It assumes uniform performance across all indicators, neglecting that a comparator may be stable for some metrics but erratic for others. It also treats comparators as statistically independent, overlooking structural correlations – such as overlap between income groups and regional classifications – which may have led to redundant weighting. Additionally, the weights are not normalised and do not sum to 1, thereby limiting their interpretability and preventing relative comparison across utilities. These shortcomings motivate the development of a more refined, indicator-specific weighting model that incorporates both local consistency and inter-comparator relationships.

7.2.2 Advanced Weights

A second approach is tested that computes indicator-specific weights based on the joint variability of z-scores across comparators. The rationale is that the stability of a comparator may differ substantially between indicators – for example, income-based grouping may be a strong benchmark for financial KPIs but poorly structured for customer service KPIs.

This relies on the use of a covariance matrix, constructed from the z-score distributions of the five comparators for each indicator. The covariance matrix captures not only how dispersed each comparator's values are, but also how their variations interrelated – a critical insight when dealing with overlapping or correlated benchmarking structures. In essence, comparators whose z-scores vary erratically or in tandem with unstable peers are weighted down, while those demonstrating unique and consistent explanatory power are weighted up.

Mathematically, the matrix $\sum_j$ is computed separately for each indicator j, where each element $\sum_{i,k}$ represents the covariance between comparators i and k. To transform this matrix into interpretable weights, each column is normalised such that the weights assigned to the five comparators for a given indicator summed to 1:

$$\text{Weight}_{i,j} = 1 - \frac{\sum_{k=1}^{5} \text{Cov}(Z_{k,j}, Z_{i,j})}{\sum_{k=1}^{5} \text{Cov}(Z_{k,j}, Z_{k,j})}$$

$$\text{and then normalised such that} \quad \sum_{i=1}^{5} \text{Weight}_{i,j} = 1$$

This yields a **15×5 matrix of weights**[89], where each row corresponds to a specific indicator and each column to a comparator. The normalised format ensures that the final composite z-scores remains comparable in scale across indicators, enabling more consistent flag severity interpretation downstream.

In Figure 7.1, and in the aggregated overview in Table 7.2, the covariance-sum weights for all 15 indicators are presented across the 5 comparators. This provides a more nuanced perspective on representativeness and comparator reliability than the original simple weighting approach.

[89] The full results are presented in Appendix A.5.

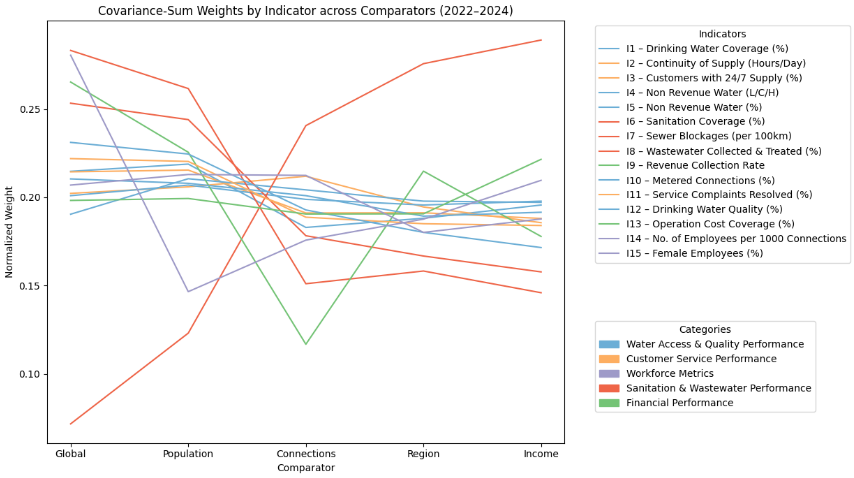


Figure 7.1: *This figure presents the covariance-sum weights for each indicator across the five comparator groups, with indicator colours aligned to the categories defined in Chapter 2.1, facilitating semantic interpretation.*

Table 7.2: *Average Advanced Weight Results across all 15 Indicators for the NewIBNET 2022-2024 Dataset*

Comparator	**Average Weight**
Global	0.216
Population	0.208
Connections	0.193
Region	0.193
Income	0.189
Total	1.000

Global carries the highest average weight (≈0.216), highlighting its relatively unique contribution across indicators. Population follows at around 0.208, still important but closer to the pack. Both Income (≈0.193) and Region (≈0.193) settle in the mid-range, indicating moderate overlap. Connections, however, remains the lowest (≈0.189), underscoring its slight redundancy with the other comparators.

Viewing the weights through the thematic categories defined in Chapter 2.1 reveals patterns in how comparator relevance varies across indicator types.

For **Water Access & Quality Performance** indicators, Population emerges as the most influential comparator, indicating that service size plays a central role in explaining variation in utility performance. **Customer Service Performance** indicators display a more balanced distribution of weights, with Population, Connections, and Global comparators contributing almost equally. This suggests that customer service dynamics are shaped by a combination of utility size, infrastructure reach, and broader systemic trends.

Workforce Metrics indicators are most strongly associated with Global and Population comparators, reflecting both structural norms and the influence of utility scale on workforce composition. For **Sanitation & Wastewater Performance** indicators, the Global comparator carries the most weight, pointing to a lack of strong regional or demographic clustering and underscoring the infrastructural variability of wastewater services. **Financial Performance** indicators are most influenced by Global and Income comparators, highlighting the importance of broader economic context and income-level effects in shaping financial performance outcomes.

These patterns show the value of the full-covariance weighting approach: by tailoring weights to each indicator's unique variability structure, it avoids the oversimplifications of uniform schemes and ensures that the most contextually appropriate comparator group is emphasised for each thematic domain.

> **Summary:** This section introduced a refined scoring approach to improve the fairness and statistical rigour of anomaly detection by assigning weights to comparator groups based on their reliability. Two weighting models were tested: a simple global variance-based method and a more advanced, indicator-specific approach using covariance matrices, which revealed nuanced differences in comparator relevance. This highlighted the Global comparator as the most statistically independent and contextually informative comparator.

7.3 Severity Flagging

With comparator weights established, this section explores an alternative threshold mechanism based on a composite severity score – a single weighted aggregation of z-score deviations across all comparators, calculated per indicator for each utility.

Unlike the current system, which evaluates each comparator-indicator pair independently, resulting in 21,675 separate z-scores in the case of 5 comparators, 15 indicators, and 289 utilities[90], this

[90] These quantities are calculated under the assumption of a complete dataset for all 289 utilities, allowing for an estimation of the upper bound of expected results.

approach condenses anomaly information into a single unified metric per indicator. This method is expected to then produce 4,335 final z-scores, as it eliminates the 5 comparator dimension, significantly reducing the information load for reviewers and enabling more scalable diagnostics.

Two areas are explored: one defining a composite severity score in Section 7.3.1, and another testing sensitivity thresholds to determine optimal cut-off points for flagging in Section 7.3.2.

7.3.1 Composite Z-Score Logic

To move beyond individual comparator flagging, this section proposes a unified thresholding method based on a composite severity score. The goal is to condense multiple comparator-based deviations into a single, interpretable metric that reflects both the magnitude and reliability of the underlying evidence. One intuitive way to achieve this is by using the previously computed weights to scale each comparator's z-score, thereby reflecting its relative diagnostic value.

However, to ensure comparability across indicators and avoid inflating scores due to weight distribution, the composite also has to be normalised. This results in a standardised weighted z-score – a continuous metric that can support thresholding, ranking, and severity tiering.

Mathematically, it is defined as a weighted linear combination of standardised z-scores:

$$Z_j^{\text{weighted}} = \frac{\sum_{i=1}^{5} \text{Weight}_{i,j} \cdot Z_{i,j}}{\sqrt{\sum_{i=1}^{5} w_i^2}}$$

Where:

- $Z_{i,j}$ is the z-score of utility u for indicator j within comparator i,
- $\text{Weight}_{i,j}$ is the pre-computed trust weight for that pair.
- The denominator $\sqrt{\sum_{i=1}^{5} w_i^2}$ serves as a normalisation factor that preserves the scale and interpretability of the resulting weighted z-score. Although the covariance-based weights for each indicator are normalised to sum to 1, this does not ensure that the magnitude of the resulting composite score is comparable to a standard z-score. Without this adjustment, the resulting score could be artificially deflated or inflated depending on how the weights are distributed[91]. The denomi-

[91] For example: one dominant weight vs. five evenly spread.

nator thus ensures that when weights are applied, the resulting severity score maintains the statistical properties of a standard deviation-based z-score.

To illustrate this process, Table 7.3 presents an example calculation for I1 using a sample utility. It displays the weights per comparator for I1, the utility's z-scores for I1 relative to each of the five comparator groups, and the intermediate weighted values prior to normalisation, with the squared weights shown in the final column.

Table 7.3: *Example Indicator 1 Calculations of a Composite Score for a Utility from the NewIBNET 2022-2024 Dataset*

Comparator	**Weight**	**Z-Score**	**Weighted (w·Z)**	**Weight^2**
Global	0.2147	−1.83	−0.3929	0.04610
Population	0.2188	−1.65	−0.3610	0.04787
Connections	0.1829	−1.67	−0.3054	0.03345
Region	0.1881	−1.83	−0.3442	0.03538
Income	0.1956	−2.00	−0.3912	0.03826
Totals	1.0000		−1.7948	0.20106

Based on this data, the final weighted z-score for Indicator 1 for this utility is calculated as follows:

$$\begin{aligned} \text{Numerator:} \quad & \sum_{i=1}^{5} w_{i,j} Z_{i,1} = -1.7948, \\ \text{Denominator:} \quad & \sqrt{\sum_{i=1}^{5} w_i^2} = 0.4484, \\ \text{Weighted Z-Score:} \quad & Z_1^{\text{weighted}} = \frac{-1.7948}{0.4484} \approx -4.00. \end{aligned}$$

When the above logic is applied across all utility-indicator combinations, it produces 3061 results in the 2022-2024 dataset. The next step is to interpret these values – for example, *does a score of -4.00 reflect abnormal behaviour that requires attention?* The following section explores appropriate threshold levels based on the current mathematical construction.

7.3.2 Sensitivity Testing

With the composite z-score standardised in previous section, its distribution could be reasonably approximated as normal – a property supported by the Central Limit Theorem[92]. This assumption enables the application of the empirical rule:

[92] **Central Limit Theorem:** This suggests that linear combinations of (approximately) normal variables tend toward normality (Montgomery et al., 2014).

- ~68% of values lie within ±1 standard deviation,
- ~95% within ±2,
- ~99.7% within ±3,
- Values beyond ±4 are exceedingly rare and typically signal anomalies or structural inconsistencies.

Building on this, a tiered severity model is developed to classify the likelihood and urgency of an anomaly based on the composite z-score:

- **Normal ($|z| \leq 2$):** Within expected statistical variation.
- **Mild ($2 < |z| \leq 3$):** Potential anomaly – merits attention, but not immediately concerning.
- **Moderate ($3 < |z| \leq 4$):** Strong deviation – likely warrants follow-up or verification.
- **Severe ($|z| > 4$):** Statistically extreme – high probability of error or misreporting.

A visualisation of the normal distribution with the corresponding z-score categories is shown in Figure 7.2.

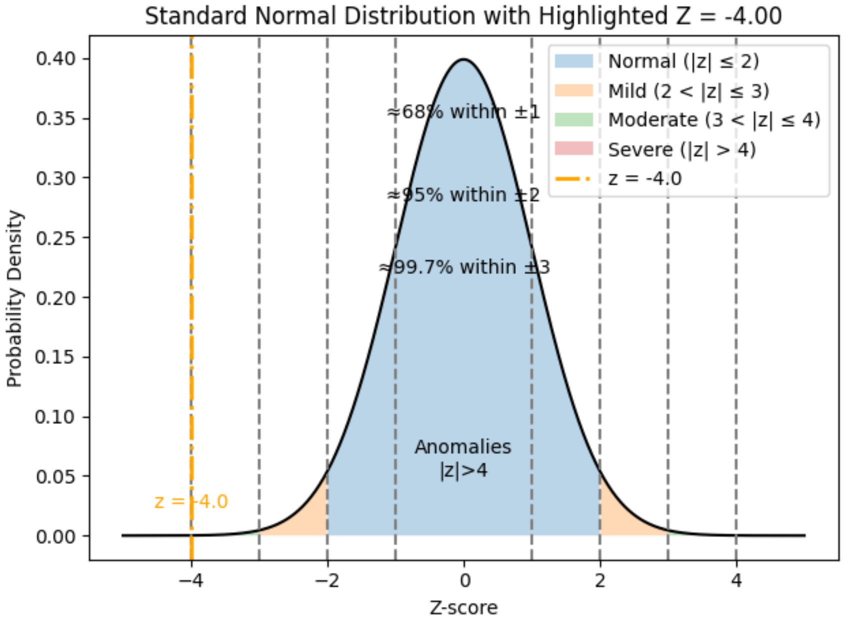


Figure 7.2: *This figure illustrates the standard normal distribution with the defined z-score categories, alongside the example from Section 7.3.1, where a composite score of approximately -4.00 is positioned within the Severe category, demonstrating how such a deviation is interpreted in terms of its statistical relevance.*

Taking the example from Section 7.3.1, the calculated z-score of -4.00 falls within the *Severe* category on the normalised distribution graph shown in Figure 7.2. This places it outside the range of normal statistical variation and classifies it as an exceedingly rare deviation, warranting follow-up or verification.

Looking at the complete results, Table 7.4 presents the number of utility-indicator combinations and their respective categories.

Table 7.4: *Severity Results for the NewIBNET 2022–2024 Dataset*

Tier	Number of Instances	Number of Utilities
Normal	2423	239
Mild	369	195
Moderate	132	102
Severe	137	87
Total	3061	

In total, 3,061 composite z-scores are generated, of which 638 are classified as non-normal deviations. The majority of cases falls into the *Normal* category, accounting for approximately 79% of all instances, while *Moderate* deviations are the least common with less than 5%.

However, a closer examination of the instances reveals that, when combining all *Mild*, *Moderate*, and *Severe* cases, 217 out of 239[93] utilities had at least one instance falling into one of these categories. This indicates that while around 20% of all utility-indicator combinations are flagged as potentially anomalous, the utility-level coverage is remarkably high at 90%.

When compared to the existing threshold logic in Section 7.1 – where only values beyond ±4 are flagged – it could be argued that only the *Severe* category aligns with this standard. Of the 137 *Severe* instances identified, 87 utilities are involved with at least one such case, significantly narrowing the scope from the earlier 217 potentially anomalous utilities out of 239. Another criterion could be assessing how many utilities had overlapping cases across all three categories. In this case, 46 utilities have at least one instance in each of the *Mild*, *Moderate*, and *Severe* categories. This raises a critical question regarding the appropriate threshold for distinguishing between genuinely anomalous behaviour and potential over-flagging.

The distribution of weighted z-scores in Table 7.5 shows a clear skew, with values ranging from -8.48 to 24.46, a median of -2.13, and an average slightly below zero.

[93] Of the 289 total utilities, 21 were flagged for incomplete inputs and 29 for rule-based logic violations in Chapter 5, leaving 239 utilities for deviation analysis.

Table 7.5: *Descriptive Metrics for Weighted Z-Score for the 2022-2024 NewIB-NET Dataset*

Metric	Value
Minimum	-8.484
Maximum	24.458
Median	-2.134
Average	-0.485

This indicates that most utilities fall below the expected benchmark, pulling the central tendency into the negative range, while a smaller number of extreme positive outliers drive the maximum far above the mean. Only a subset of cases deviate meaningfully from the benchmark, with the majority clustered closer to the *Mild* range. This suggests that while deviations exist, they are concentrated in a limited number of high-severity outliers rather than being widespread across the dataset.

In Figure 7.3, an analysis of the severity distribution across indicators in the comparator-based validation stage also reveals meaningful variation in how different metrics deviate from expected norms.

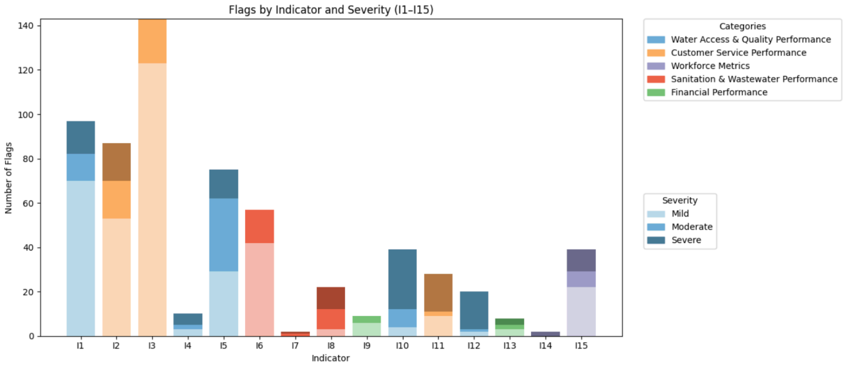


Figure 7.3: *This figure presents the distribution of identified flags by indicator and severity level for the 2022–2024 dataset, with mild, moderate, and severe bars coloured according to the categories defined in Chapter 2.1, enabling visual comparison of anomaly intensity across indicators.*

Examining the 2022–2024 flag distribution reveals that indicators drinking water coverage (I1), continuity of supply (I2), and especially customers with 24/7 supply (I3) account for the highest overall number of deviations, with I3 alone generating 143 flags – predominantly *Mild* in severity. When viewed through the thematic categories defined in Chapter 2.1, the greatest concentration of anomalies occur in **Customer Service Performance** and **Water Access & Quality Performance** indicators, suggesting that these domains are potentially prone to measurement challenges, reporting inconsistencies, or genuine operational issues.

In contrast, indicators such as metered connections (I10), service complaints resolved (I11), and drinking water quality (I12) display a high share of *Severe* deviations relative to their total counts, signalling sharper outliers and potentially more critical data quality or performance concerns. **Financial Performance** indicators show the lowest deviation rates overall, indicating a relatively stable reporting landscape in that dimension.

The severity breakdown shows the value of disaggregating anomalies not only by indicator but also by thematic category. This approach enables targeted, context-aware validation logic that can inform both refinement and follow-up investigations – whether to address systemic measurement errors, outdated equipment, or persistent inconsistencies in reporting practices.

To conclude, this method replaces ad hoc flagging rules with a theoretically grounded framework that scales severity based on statistical extremity. Crucially, the comparator weights introduced earlier refine this signal: comparators with higher internal variance contributed less to the final score, reducing noise from inconsistent groupings. As a result, high composite z-scores reflect both the magnitude of deviation and the robustness of supporting evidence.

> **Summary:** This section introduced a composite severity scoring method that condenses multiple comparator-based z-scores into a single, weighted metric per indicator, streamlining the anomaly review process. By applying normalisation and probabilistic thresholds, the method enabled tiered severity classification (*Normal*, *Mild*, *Moderate*, *Severe*), revealing that although 20% of data points were flagged as anomalies, 217 out of 239 utilities had at least one flagged indicator, raising important questions about over-flagging versus genuine abnormality in the dataset.

8
Technical Validation & Evaluation

> **RQ4:** *To what extent does the proposed anomaly flagging system perform reliably, and align with expert validation and benchmarking expectations?*

This chapter brings together the analytical insights and experimental findings from the preceding chapters to offer a system-level perspective on the proposed validation framework. Moving beyond individual components, it aims to evaluate how these elements interact as part of a coherent pipeline. The chapter begins with an architectural synthesis in Section 8.1, translating the developed anomaly detection stages into a structured, reviewer-facing system designed for practical implementation within NewIBNET. This is followed by internal system testing in Section 8.2, which probes the pipeline's robustness, limitations, and performance when applied to test datasets, providing a proxy for 'ground truth' evaluation. Finally, the chapter closes with expert feedback integration in Section 8.3, recognising that a system's effectiveness is not only technical but also social: stakeholder trust and interpretability are critical for successful real-world adoption.

8.1 Architectural Synthesis

The preceding chapters – Chapter 5: Data Preparation & Structural Validation, Chapter 6: Context-Aware Anomaly Modelling Frameworks, and Chapter 7: Severity Scoring & Decision Framework –

collectively laid the foundations for a comprehensive data validation pipeline.

What began as discrete validation experiments evolves into a structured, multi-stage pipeline. This evolution is not predefined, but emerges logically from empirical observations, literature-backed design choices, and iterative refinement across data preparation, contextual comparison, and severity modelling.

Section 8.1.1 offers a deeper dive into how the 2022–2024 utility dataset is processed across each validation stage, quantifying the impact and distribution of flags throughout the pipeline. Section 8.1.2 then consolidates the full methodology into a proposed end-to-end validation system, complete with defined flag categories, threshold logic, and architectural flow, offering a tangible, operational prototype for future application within the NewIBNET framework.

8.1.1 Cumulative Flag Analysis

The architecture that emerges from a series of research experiments is divided into three core stages:

- **Stage 1 (Structural Validation):** The pipeline begins by screening for foundational data quality issues. Utilities are flagged for missing values, placeholder entries, and type errors that compromise the interpretability or validity of the dataset.
- **Stage 2 (Indicator Logic):** Automatic logical validation rules are applied at the indicator level. This includes coherence checks between raw and derived values, as well as internal consistency validations across related indicators. Utilities that fail these domain-informed checks are flagged for further review.
- **Stage 3 (Comparator Analysis and Severity Scoring):** For utilities passing the previous stages, final anomaly detection is performed using comparator-based analysis. Five context-aware comparator groupings are formed to generate z-scores per utility-indicator pair. Recognising that these comparators vary in statistical robustness, a weighted scoring model is applied to compute composite deviation scores. These scores are then evaluated against probabilistic thresholds, enabling tiered severity flagging for outlier detection based on statistical magnitude and comparator reliability.

The layered nature of the framework allows for granular tracking of where and why utilities are flagged throughout the validation process. As shown in Figure 8.1, each stage introduced flagging checks.

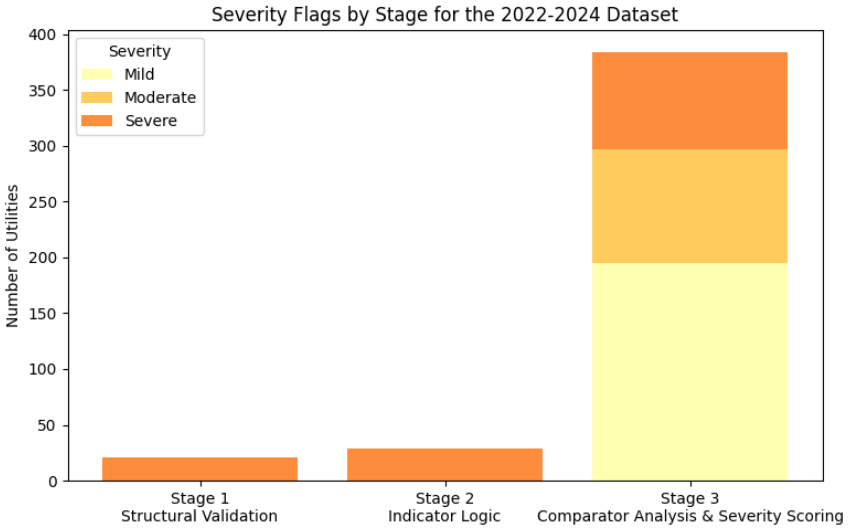


Figure 8.1: *This figure shows the severity flags by stage for the 2022–2024 dataset. Since Stage 3 (Comparator Analysis and Severity Scoring) is the only stage that distinguishes between different severity levels, three separate colours are used.*

Across the three validation stages, the automated pipeline flags a substantial proportion of the dataset. In **Stage 1 (Structural Validation)**, 21 utilities are identified with structural issues. In **Stage 2 (Indicator Logic)**, 29 additional utilities are flagged for violating indicator-level logic. **Stage 3 (Comparator Analysis and Severity Scoring)** flags 217 utilities with at least one indicator falling into the *Mild*, *Moderate*, or *Severe* categories. This segmentation in **Stage 3 (Comparator Analysis and Severity Scoring)** highlights an important interpretative tension: *how should severity be operationalised in the review process, and what constitutes a true anomaly in the absence of ground truth?*

As this marks the first large-scale application of automated scoring within NewIBNET, it remains plausible that many of these flags reflect genuine issues previously undetected through manual review. The results suggest that up to 267 out of 289 utilities (92%) may require further attention if all flagged cases across the three stages are considered. Even under a conservative lens, focusing solely on *Severe* cases across all stages, 137 utilities still merit manual review. While this figure may appear overwhelming, it is important to note that each flagged instance is now traceable to a specific rationale. This traceability provides structure to the review process, potentially reducing time burden and increasing diagnostic clarity for expert reviewers.

Together, these stages represent a pipeline grounded in technical validity.

8.1.2 Flag Pipeline

It is established that anomalies in the NewIBNET dataset emerge not from a single source, but through a layered interplay of issues. These distinct sources of error can be formalised into a three-stage validation process. Collectively, they inform the definition of what this system considers a flagged entry.

> In this context, a *flag* is defined as any data point that:
>
> - is incomplete where completeness is expected,
> - has an incorrect data type,
> - contains a placeholder value (e.g., ’1.0’),
> - violates indicator logic, or
> - deviates substantially from established comparator group norms

Drawing inspiration from dual-layer metadata validation systems (Y. Liu et al., 2025), the proposed architecture seen in Figure 8.2 incorporates two validation layers to identify flags in accordance with the aforementioned definition: a front-end responsible for structural and logical checks, and a back-end that handles contextual benchmarking and severity scoring.

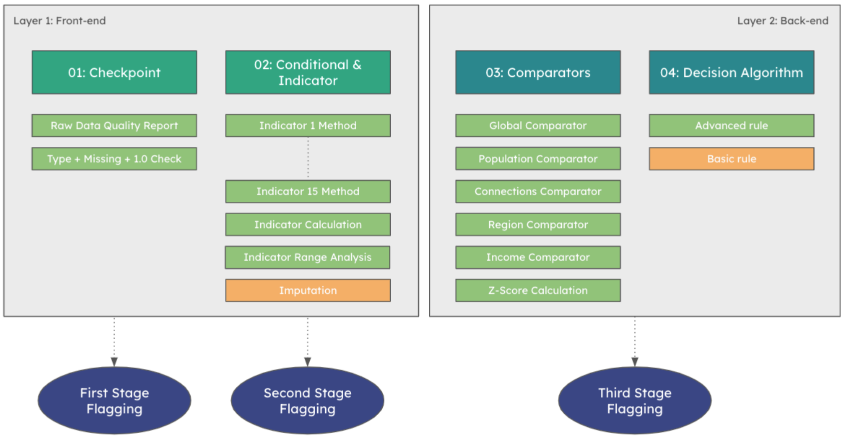


Figure 8.2: *This diagram illustrates the front-end and back-end components of the newly proposed data review pipeline, highlighting the three key stages where automated flagging mechanisms will be implemented. The light green segments indicate core elements of the system while the orange segments indicate optional or alternative methods.*

In summary, this dual-layer validation pipeline transforms experimental findings into a deployable, auditable framework[94] for data quality assurance.

> **Summary:** This section presented the synthesis of previous research chapters into a structured, three-stage validation pipeline, covering structural checks, logical validation, and comparator-based anomaly detection, which flagged a significant portion of the 2022–2024 NewIBNET dataset. The resulting dual-layer architecture formalises these stages into a practical, auditable system for data quality assurance, offering both transparency and scalability for future implementation.

8.2 Internal System Testing

Before deploying an automated anomaly flagging system in a live benchmarking environment, it is essential to verify how it behaves under controlled but realistic conditions. Since no quantitative baseline from other benchmarking systems[95] is publicly available, direct comparison is not possible; instead, this section introduces internal testing as a structured approach to assess robustness, reliability, and efficiency. To the best of current public knowledge, this represents the first systematic attempt to evaluate such a system, addressing a notable gap in the benchmarking literature and practice.

Internal testing is divided into two main parts. First, Robustness Checks in Section 8.2.1 explore how the system responds to different scenarios. These checks aim to identify any blind spots, inconsistencies, or unnecessary computational burdens that could undermine system performance.

Second, the Case Study Application in Section 8.2.2 applies the system to a set of real-world utility profiles from Indonesia, allowing for a closer examination of how the flagging logic operates in a different context.

8.2.1 Robustness Checks

This section evaluates the impact of imputation on flagging results, the detection of test inputs within the 2022–2024 dataset, and provides a general overview of the current computational efficiency

[94] For more information on how this framework is developed and shared with NewIBNET for practical use, see Appendix D.2.

[95] Other benchmarking platforms reviewed in Appendix B.1, such as AWWA and EBC, reported no use of automated flagging systems.

and performance of the pipeline.

Impact of Imputation

Building on the imputation approaches outlined in Chapter 5.3 – where 9 sporadic[96] wastewater indicators are first flagged and partially complete entries imputed using k-NN or median substitution – Figure 8.3 shows how these strategies affect severity outcomes when applied through the full pipeline.

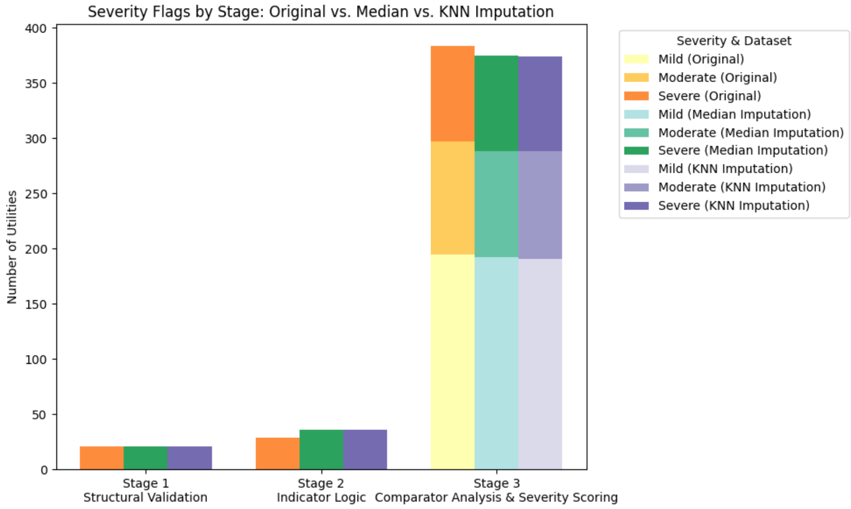


Figure 8.3: *This figure presents the distribution of severity flags by stage for the original NewIBNET 2022–2024 dataset, compared against versions processed using median imputation and k-NN imputation.*

The impact is relatively modest, but consistent: both imputation methods slightly reduce the total number of utilities and severity instances, with the largest relative changes seen in the *Moderate* and *Severe* tiers. This suggests that imputing missing wastewater values helps smooth out anomalies that arise from incomplete records, reducing noise without significantly altering the overall distribution of flags.

The differences between k-NN and median imputation are marginal, with k-NN showing a slightly greater reduction in *Moderate* flags, while *Severe* flags remain largely stable.

In practice, this indicates that robust handling of partial wastewater data can modestly enhance statistical stability in comparator-based flagging without introducing major shifts in severity classification.

[96] Chapter 5.3 provides a detailed overview of the categorisation.

Detection of Test Inputs

Test inputs are defined as entries submitted either through the official online survey form or via Excel by World Bank staff for the purpose of testing the IT functionality of the system, rather than evaluating indicator performance. These entries are not grounded in real utility data and are typically arbitrary, often containing implausible values. As noted in Chapter 5, the discovery of a "utility" reporting a service population of 2 billion exemplifies the type of unrealistic placeholder data that can appear in the dataset. If left undetected, such entries risk being treated as genuine submissions, particularly in **Stage 3** (**Comparator Analysis and Severity Scoring**), where they could introduce unnecessary variance and undermine the robustness of comparator-based analysis. While the current covariance-based weighting smooths much of this noise, it cannot fully eliminate its impact, indicating the importance of test input detection.

The results of test input detection for the 2022–2024 dataset are shown in Figure 8.4.

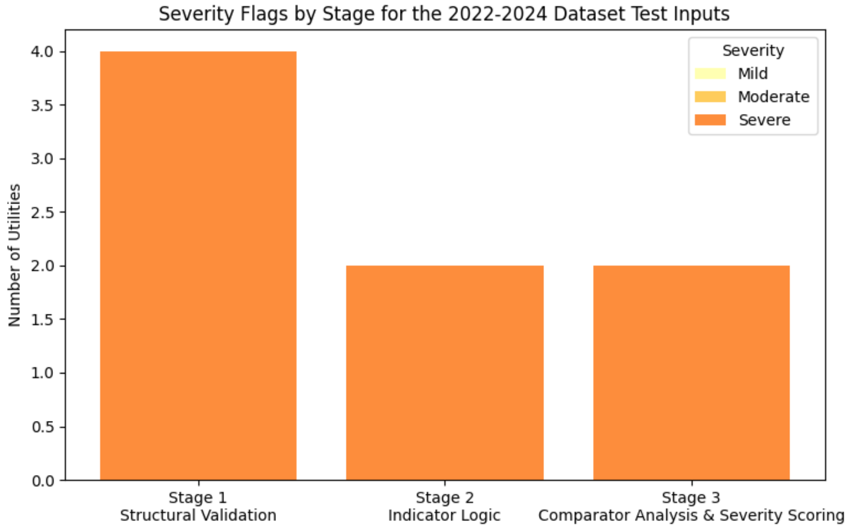


Figure 8.4: *This figure shows the severity flags by stage for the 2022–2024 dataset test inputs. Since Stage 3 is the only stage that distinguishes between different severity levels, three separate colours are used.*

8 such test entries are identified based on explicit naming[97]. 4 are captured immediately in **Stage 1** (**Structural Validation**), 2 in **Stage 2** (**Indicator Logic**), and 2 persisted until **Stage 3** (**Comparator Analysis and Severity Scoring**). While all 8 are ultimately flagged, the fact that a quarter reached the final stage shows the importance of systematically identifying and filtering test inputs early to preserve dataset integrity.

[97] Test inputs are identified by utility names containing the keyword "test". Since names are not used in flagging decisions, this provides a reliable method for detecting them.

Computational Efficiency & Performance

Assessing the computational efficiency of the automated flagging pipeline is essential to ensure that scalability is maintained as the dataset grows, avoiding excessive runtimes or system instability. Table 8.1 summarises the runtime and cyclomatic complexity[98] for each processing stage in a single run of the complete system.

Table 8.1: *Runtime and Cyclomatic Complexity per Processing Stage*

File	Runtime (s)	Cyclomatic Complexity
Cleaning	< 1.0	29
Stage 1	7.6	15
Stage 2	10.0	82
Stage 3	56.9	88

Runtime captures the relative processing cost of each stage in seconds, while cyclomatic complexity reflects code maintainability by measuring the number of distinct decision paths. Although complexity is not a direct indicator of performance, high values signal substantial branching and conditional logic that can complicate testing and increase the likelihood of edge-case errors.

Profiling results show **Stage 3 (Comparator Analysis and Severity Scoring)** as the dominant bottleneck (≈57 seconds), followed by **Stage 2 (Indicator Logic)** (≈10 seconds), with **Cleaning** and **Stage 1 (Structural Validation)** contributing negligibly. This distribution stems from the computational intensity of multi-group z-score calculation, comparator aggregation, and weighting in **Stage 3 (Comparator Analysis and Severity Scoring)**, and the rule-based indicator checks in **Stage 2 (Indicator Logic)**.

The pipeline currently scales linearly with dataset size, but substantial efficiency gains[99] are achievable without altering outputs.

This offers an overview of the proposed pipeline's complexity and efficiency: reducing 75 hours of manual checks to under 2 minutes. The key question, however, is whether it matches the effectiveness and accuracy of manual review in determining flags.

8.2.2 Case Study Application in Indonesia

As part of the internal validation process, it is important to apply the pipeline to a different dataset of utilities to explore whether discernible trends, differences, or similarities can be related back to the research findings or motivate potential modifications to the pipeline. This exercise is particularly valuable given that the the-

[98] **Cyclomatic Complexity:** A software metric that counts the number of linearly independent execution paths through the code. Higher values indicate more branching, which can challenge maintainability and testing (McCabe, 1976).

[99] Detailed suggestions for improving performance efficiency are provided in Appendix B.2.

sis has thus far examined the system from only a single perspective. The external dataset, comprising 398 records from 2024, is provided by PERPAMSI[100], the Indonesian Water Supply Association, and covers only utilities operating in Indonesia.

This yields different dynamics compared to the full NewIBNET 2022–2024 dataset, as illustrated in Figure 8.5. Only 14 of the original 24 raw input columns could be mapped[101], enabling the calculation of 8 out of 15 indicators. This reduction in indicator coverage inherently narrows the scope of structural and logical checks.

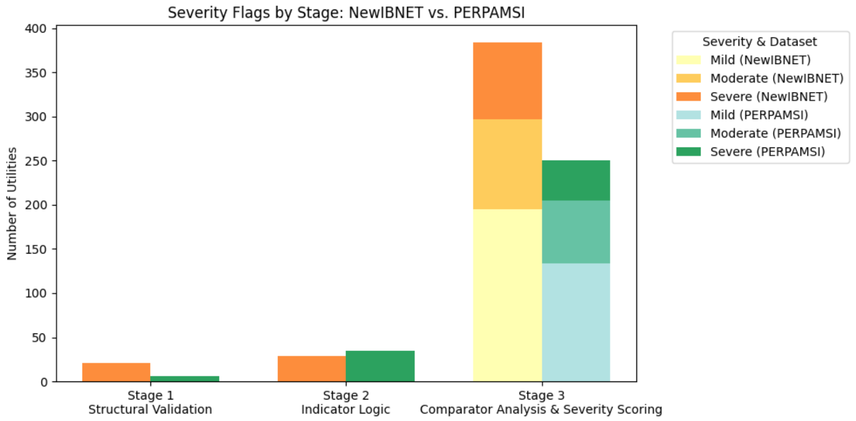


Figure 8.5: *This figure compares the distribution of severity flags across stages for the PERPAMSI 2024 dataset and the NewIBNET 2022-2024 dataset.*

Despite these constraints, the proportion of flagged utilities remained substantial: 207 out of 357 utilities (≈58%) receive at least one type of flag in **Stage 3 (Comparator Analysis and Severity Scoring)**, with a distribution skewed towards mild anomalies. This is broadly consistent with the relative scale of **Stage 3 (Comparator Analysis and Severity Scoring** outputs in the full NewIBNET dataset, though here the smaller indicator set and national scope mean that anomalies are identified within a more homogenous operational and contextual environment. Notably, **Stage 2 (Indicator Logic)** flags are concentrated entirely on population service size reporting, with 35 cases where the reported population served exceeds the total population – an implausible scenario suggesting a single dominant reporting error rather than the diversified spread across indicators seen in the NewIBNET dataset.

At the indicator level, the PERPAMSI dataset reveals a similar profile compared to the NewIBNET sample. For instance, operational cost coverage (I13) remains largely within the 100–150% median, but extreme outliers above 1,000% persist, suggesting the presence of classification errors even in a more homogeneous national dataset.

[100] **PERPAMSI** was founded in 1972 and serves as the national network of regional drinking water utilities (PDAMs) in Indonesia. Read more: https://www.perpamsi.or.id/

[101] The full mapping and detailed results for the PERPAMSI dataset are provided in Appendix C.1.

The recalculated covariance-based weights, excluding region and income comparators[102], shift emphasis strongly towards the Population comparator group, with weights frequently exceeding 0.45. This reflects reduced cross-comparator variability in the absence of regional and income dimensions, thereby amplifying the influence of population-based benchmarking. The Population and Connections category distributions also reveal a high concentration of *Small* utilities, potentially biasing comparator statistics and severity thresholds towards the operational realities of lower-capacity providers.

The similarity in overall flagging proportions to the full NewIBNET dataset, despite the reduced indicator set and national scope, suggests two possibilities for **Stage 3** (**Comparator Analysis and Severity Scoring**). It may reflect a robust framework that maintains detection rates across contexts, or it may be a by-product of heavy weighting towards the Population comparator and a skewed utility size distribution, which tightens variance bounds and risks over-flagging. Given that the system is inherently dependent on the data it receives – without reference to external industry benchmarks – any widespread deviation from sector standards will shift internal baselines, meaning that "anomalies" may simply reflect divergence from a non-representative norm. This reinforces the need for sensitivity analysis[103] to distinguish genuine performance issues from artefacts of comparator homogeneity.

> **Summary:** This section evaluates the proposed flagging system under controlled conditions through robustness checks and a real-world case study with Indonesian utility data. The results highlight how imputation, test input detection, computational efficiency, and contextual dataset differences influence flagging outcomes, offering insights into both the system's stability and areas for refinement before wider deployment.

8.3 Expert Feedback

To ensure the practical relevance and usability of the proposed flagging system, it is essential to gather input from those most familiar with the operational realities of utility data validation. While the preceding chapters developed and evaluated the system through

[102] The PERPAMSI dataset covers only Indonesian utilities, meaning all belong to the same region and income group, rendering these comparators redundant in the analysis.

[103] **Sensitivity Analysis:** Tests how robust flagging results are by adjusting comparator weights and thresholds to see if anomalies persist or stem from statistical artefacts.

technical experimentation, the transition from a theoretical model to an implementable tool demands validation by domain experts. This section focuses on integrating that practitioner perspective.

An anonymous online survey is conducted with the NewIBNET team. Their insights offer critical reflections on the design logic, perceived usefulness, and potential challenges of implementing the proposed architecture in practice. The section is structured as follows: Survey Design in Section 8.3.1 outlines the structure, rationale, and content of the survey; Evaluation Insights in Section 8.3.2 synthesises the feedback received, presenting key findings, selected quotations, and a thematic analysis of expert perspectives. These inputs inform the final considerations of system adoption, user trust, and institutional feasibility.

8.3.1 Survey Design

To evaluate the practical applicability and perceived value of the proposed automated flagging system, an anonymous online survey[104] is conducted among experts directly involved with the NewIBNET system. Unlike the small group of technical specialists consulted during the preparatory phase, this survey targets a broader set of practitioners within the NewIBNET team who possess direct experience with the system's architecture and workflows. The goal is to collect both qualitative and quantitative feedback on the system's decision logic, usability, and potential integration into real-world workflows.

The full survey is provided in Appendix C.2. For a general overview, it comprises three main sections:

1. **Case-Based Evaluation:** Participants review two fictional but realistic utility profiles containing common data anomalies[105]. They are asked whether they would flag each utility and why. Afterwards, the same cases are shown with the system's flagging results, allowing participants to reflect on its logic and accuracy. In short, Utility A (Singapore) tests expert reactions to critical missing values and wastewater service data, while Utility B (The Netherlands) presents a potential financial anomaly that bypasses **Stage 2** (**Indicator Logic**) thresholds yet may still warrant flagging.
2. **Flag Interpretation & Trust:** This section introduces the full validation pipeline and asks experts to evaluate the system's transparency, interpretability, and the usefulness of its severity tiers in supporting manual review.

[104] This survey was approved according to the Delft University of Technology Human Research Ethics guidelines. Refer to: `https://www.tudelft.nl/over-tu-delft/strategie/integriteitsbeleid/human-research-ethics`

[105] For example: missing values, placeholders, and implausible figures.

3. **Future Directions & Use:** Participants provide open feedback on potential system extensions, including use cases beyond anomaly detection.

The survey aims to assess not only technical performance but also expert trust, interpretability, and institutional readiness.

8.3.2 Evaluation Insights

Overall, the feedback reveals a strong endorsement of the system's foundational logic, particularly in identifying clear anomalies such as implausible placeholder values or structural inconsistencies – as seen in the unanimous agreement among all five respondents to flag Utility A (Singapore) in Figure 8.6. By contrast, views on Utility B (The Netherlands) varied, with one respondent disagreeing with the majority. While the sample size is limited, the fact that even a small group produces divergent interpretations highlights the importance of integrating deeper financial context or threshold-based checks, particularly for high-income countries where anomalies may be less clear-cut.

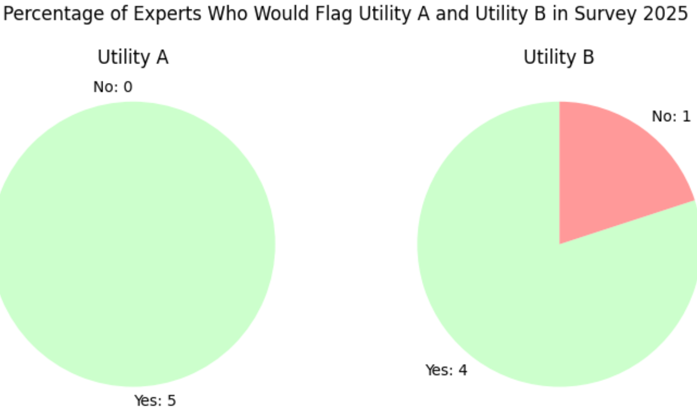


Figure 8.6: *This figure shows the number of experts in the 2025 survey who would flag Utility A (Singapore) and Utility B (The Netherlands). Green represents the share of respondents who would flag the utility, while red represents those who would not.*

When shown the system's internal imputation strategy for missing wastewater values, expert opinions are divided – 2 are in favour and 3 opposed, as illustrated in Figure 8.7. While some experts see it as a practical and necessary step to maintain analytical continuity, others stress the risk of misinterpretation[106], especially when a utility legitimately lacks a wastewater component. The concern is less about the act of imputation itself and more about how it is communicated. Several experts explicitly suggest improving the

[106] As one expert noted: *"Many utilities do not provide wastewater services but only freshwater services. To not provide them is completely reasonable."*

phrasing of imputed values and ensuring that the absence of services is recognised as a valid, meaningful data point rather than a gap to be filled. These reflections suggest that any form of algorithmic estimation must be paired with transparent language and clear visual cues.

Severity scoring, by contrast, is broadly well received, as shown in Figure 8.7. Experts appreciate its potential to prioritise attention, triage issues efficiently, and even serve as a 'data quality thermometer' – not necessarily for external users, but as an internal tool for guiding review workflows. A recurring suggestion is to refine the vocabulary of the severity labels, making them more intuitive for less technical audiences without losing the underlying statistical rigour.

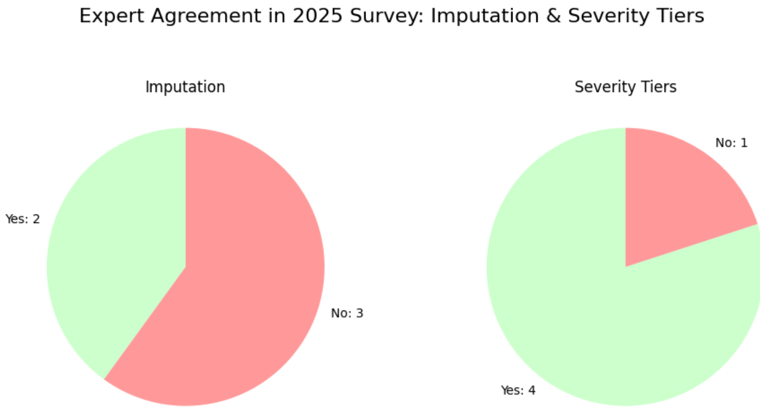


Figure 8.7: *This figure presents the results of the 2025 expert survey on the imputation and severity tier questions, with green indicating agreement and red indicating disagreement.*

Finally, the most forward-looking insight comes from how experts envision the system being used beyond flagging anomalies. Suggestions – displayed in Figure 8.8 – range from benchmarking dashboards to regional performance alerts and capacity-building tools, indicating a strong appetite for integrating this system into broader analytical and operational workflows. There is also enthusiasm for expanding its functionality to include historical consistency checks and integration with external data sources, such as annual utility reports, to enhance its diagnostic capabilities.

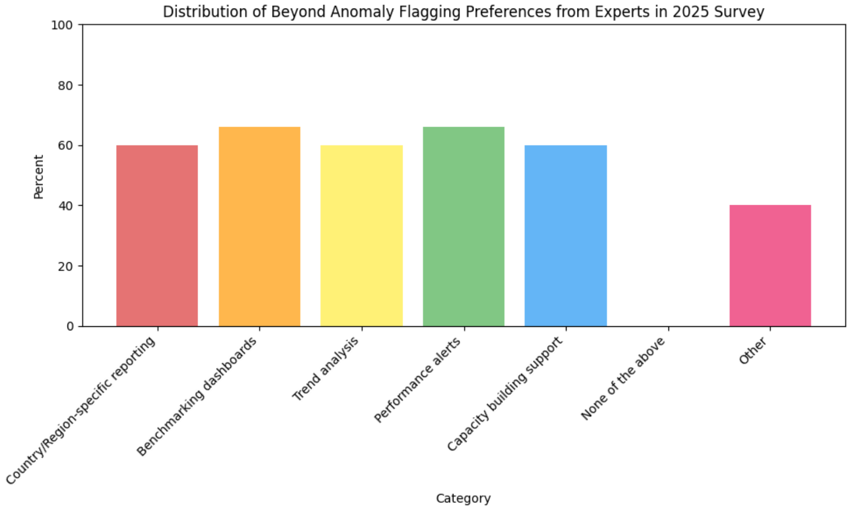


Figure 8.8: *This figure shows the distribution of responses from the 2025 survey on possible routes beyond anomaly flagging, with each bar representing a different option.*

These responses not only validate the proposed system design but also offer a user-informed roadmap for future improvements. Experts emphasise the importance of communicative clarity, adaptive thresholds, and modular flexibility – key components for a tool intended to support meaningful action in real-world benchmarking and utility engagement. Most notably, 100% of experts indicate they would use the automated flagging system, indicating both its practical relevance and perceived value.

> **Summary:** Expert feedback strongly supports the system's core logic, particularly for detecting clear anomalies, while highlighting areas for refinement such as contextual thresholds, imputation communication, and user-friendly severity labels. Respondents also envision applications beyond anomaly detection, with unanimous agreement on the system's usefulness and clear priorities for enhancing clarity, adaptability, and integration into broader benchmarking workflows.

9
Ethical, Political, and Sectoral Considerations

> **RQ5:** *What ethical and institutional implications arise from implementing an automated anomaly detection framework in the context of global water utility benchmarking?*

This chapter critically examines the ethical and institutional implications of implementing an automated anomaly detection framework in the context of global water utility benchmarking. It is organised around three interrelated themes: design reflections in Section 9.1, framing sensitivities in Section 9.2, and trust in automated flagging in Section 9.3.

These discussions address the sub-question and analyse how such a framework can be designed to remain transparent, inclusive, and ethically responsible. In doing so, the analysis is mapped against recognised standards such as the EU AI Act's risk-based categories[107] and the IEEE Ethically Aligned Design principles[108], signalling that safeguards are not ad hoc but grounded in established ethical frameworks. Moreover, the discussion recognises that IBNET data has historically informed policy reports and comparative analyses (Manghee et al., 2012; C. v. d. Berg et al., 2017; Andrés et al., 2020; Detroz et al., 2017; Tsagarakis, 2018), giving it tangible empirical impact. This creates an ethical responsibility to ensure that outputs from NewIBNET remain representative and trustworthy, as the advice and evidence derived from it may shape real-world decisions.

[107] The **EU AI Act** (**2024**) classifies AI systems by risk level, with critical water infrastructure contexts falling under 'high-risk' obligations. Read more: https://artificialintelligenceact.eu/high-level-summary/

[108] The **IEEE Ethically Aligned Design** framework outlines principles such as transparency, accountability, and human well-being in AI systems. Read more: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

9.1 Design Reflections

This section reflects on three critical design considerations that emerged during system development: dependency on utility-submitted data in Section 9.1.1, imputation and inclusivity in Section 9.1.2, and the challenge of accounting for heterogeneity and contextual realities in benchmarking in Section 9.1.3. Together, these reflections highlight the need for a benchmarking framework that is both technically robust and sensitive to the realities of the global water sector.

9.1.1 Dependency on Utility Data

A core limitation of the current framework is its reliance on submitted datasets as the sole reference point for anomaly detection, without access to a definitive *ground truth* or universally accepted industry benchmarks. This creates a self-referential baseline: deviations in **Stage 3** (**Comparator Analysis and Severity Scoring**) are measured against other submissions, not against independently verified sector standards. In contexts where performance figures inform funding allocations, regulatory scrutiny, or international comparison, there is a political dimension to whether numbers are understated, overstated, or selectively reported. Blindly trusting raw submissions risks embedding unrepresentative baselines into the benchmarking system and may unintentionally reward strategic self-presentation over transparency.

As an exploratory exercise, Benford's Law[109] (Benford, 1938) – a statistical pattern describing the expected frequency of leading digits in naturally occurring datasets – is applied to all raw inputs of the 2022-2024 NewIBNET dataset as seen in Figure 9.1. The process involves extracting the first significant digit of each raw input value, calculating its frequency distribution, and comparing it to the Benford distribution using a chi-square test. A p-value above 0.05 is interpreted as conformity.

[109] *"Examples of Benford's law being applied to fraud detection abound, from Greece manipulating macroeconomic data in its application to join the eurozone to vote rigging in Iran's 2009 presidential election."* (Murtagh, 2023)

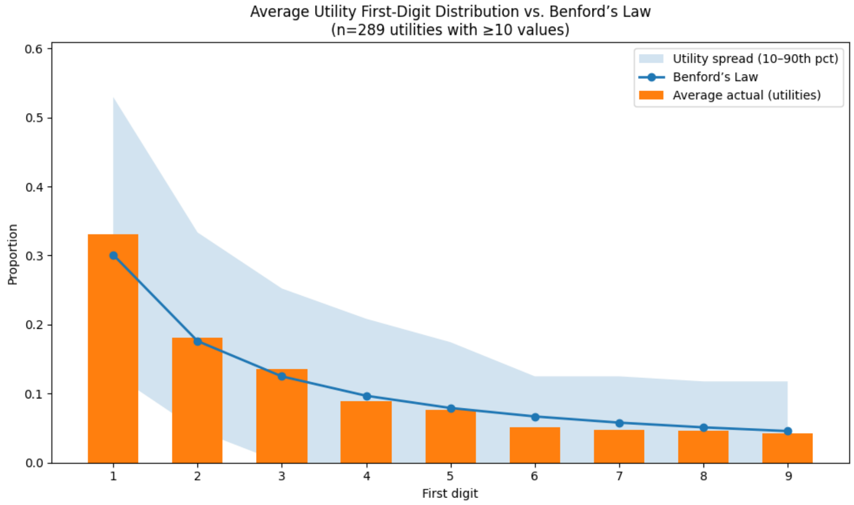


Figure 9.1: *This figure shows the average first-digit proportion distribution of utilities in the 2022–2024 NewIBNET dataset (with more than 10 raw inputs), overlaid with Benford's curve.*

The graph shows that utility data broadly follows Benford's Law: average proportion of leading digits 1–5 align closely with expectations, while digits 6–9 display systematically lower prevalence. Variability across utilities is evident, with especially wide spreads for digits 1 and 2. Detailed results show that 75 utilities do not conform. The method correctly identifies 7 of 8 test cases containing artificial values. While this does not prove manipulation – deviations may stem from sector-specific operational realities or reporting formats – it illustrates how statistical irregularity tests can complement existing plausibility checks by highlighting entries warranting closer review. A detailed overview of the Benford's Law application is provided in Appendix B.3.

From this perspective, anomaly detection is not only a statistical challenge but also a governance one. Statistical safeguards such as Benford's Law can help identify implausible values, but they cannot replace the need for sector expertise in interpreting anomalies. Ultimately, a data-driven system in a complex sector like water can never be fully complete from a purely computational standpoint. Its credibility depends on the continuous interplay between automated detection, expert oversight, and an evolving understanding of what constitutes meaningful and fair comparison.

9.1.2 Imputation & Inclusivity

In global systems such as NewIBNET, participation is not only reliant on a utility's willingness to submit data but also on its *ability* to provide complete and non-deviating values. This creates an inherent tension: enforcing completeness safeguards database integrity, yet risks excluding utilities whose operational or resource constraints limit their capacity to report certain indicators (World Bank, 2011). The wastewater dataset illustrates this dilemma. While completeness could be mandated, an expert suggestion[110] to link wastewater questions to a preliminary service-provision question – skipping them where the service is absent – would avoid partial submissions but potentially narrow participation.

Imputation offers one possible compromise. Median and k-NN approaches (Clifton et al., 2022; Miao et al., 2024) were applied to fill missing wastewater values, enabling the retention of partially complete submissions. However, the effect was limited: only 32 values were imputed, while hundreds of entirely missing cases persisted. As demonstrated in Chapter 8.2.1, the impact on comparator distributions was minimal, raising questions about whether imputation meaningfully enhances statistical robustness in its current form.

Inclusivity and data integrity cannot be treated as mutually exclusive; both require deliberate design choices that balance technical robustness with fairness in participation.

9.1.3 Heterogeneity and Contextual Realities

Representativeness plays a decisive role in how thresholds function within anomaly detection. Transitioning from a rigid binary cut-off[111] to a tiered severity scale enhances interpretability, yet thresholds rooted in classical statistical theory still rest on distributional assumptions that may fail to capture the genuine heterogeneity of utility performance. In the 2022–2024 NewIBNET dataset, 217 of 239 utilities are flagged – a result that may reflect systemic variation in performance, but could equally signal over-sensitivity to natural diversity. The 2024 PERPAMSI case study in Chapter 8.2.1 revealed a similar spread, suggesting that wide variation in plausible values is intrinsic to the sector and that overly strict cut-offs risk over-identification.

These design choices carry tangible consequences. In a sector where benchmarking influences funding, policy decisions, and leadership stability, a flag is not a neutral signal (World Bank, 2011): it can trigger political scrutiny and compel utilities to defend performance that may, in fact, be due to structural constraints rather than operational failings. Conversely, the absence of a flag may allow persistent underperformance to escape detection. From an eth-

[110] *"The other option would be to link this question with the question where it is asked if the utility provides wastewater services/sanitation and if the question is no, then automatically the question on sanitation will not come up in the questionnaire."*

[111] A utility is flagged if the z-score for any indicator fell outside the range $[-4, 4]$.

ical perspective, the decision framework functions as a gatekeeper, shaping who is scrutinised, on what grounds, and with what potential repercussions.

A further limitation lies in the framework's reliance on predefined static comparators – such as Population and World Bank Income Level – which, while enhancing interpretability and aligning with established reporting norms, can obscure meaningful intra-group variation. As demonstrated in comparative policy research (Riley et al., 2021), such aggregation may mask critical differences in infrastructure quality, regulatory capacity, climate risk, and post-conflict recovery conditions, thereby constraining the contextual accuracy of cross-group comparisons. Without incorporating macro-level or dynamic comparators[112] the framework risks reinforcing biases that penalise utilities operating in the most challenging environments.

Following findings from fairness-in-algorithms literature (Sarkar, 2022), distributional analyses are essential to ensure that anomaly detection outcomes are not systematically influenced by comparator definitions, underlying data imbalances, or contextual factors that may disadvantage certain utility groups. The analysis in Figure 9.2 reveals the flagging patterns[113] across comparator groups.

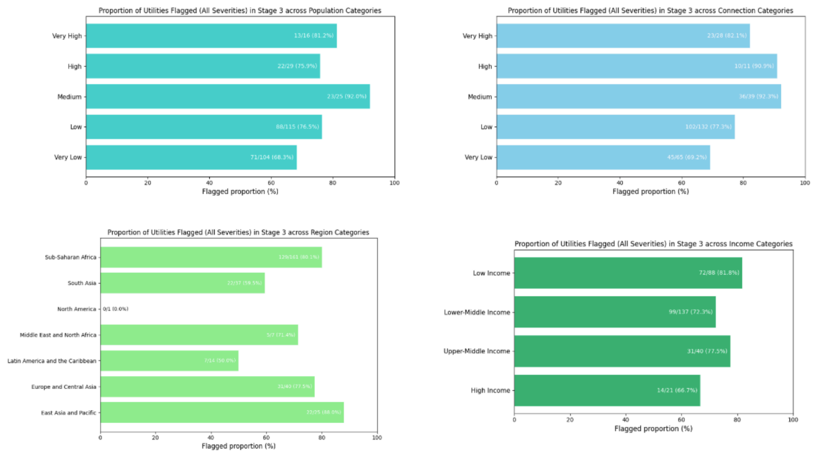


Figure 9.2: *This figure shows the proportion of utilities flagged across all categories and comparators.*

The results show that smaller groups, such as *Medium*-sized utilities or certain regional categories, often display disproportionately high flagging rates, likely due to sample size effects rather than systematic data issues. Conversely, larger categories like *Low* and *Very Low* populations account for the majority of flags in absolute terms. Regional variation further indicates the role of group

[112] Macro-level or dynamic comparators could incorporate factors such as conflict exposure, fiscal volatility, climate shocks, and other relevant dimensions not yet addressed in the current pipeline.

[113] A more detailed analysis of flagging proportions by group and comparator is provided in Appendix C.3.

size in shaping outcomes, while the higher proportions observed among *Low*-Income utilities point to possible capacity or reporting constraints. Overall, these patterns highlight the need to interpret flagging rates with caution, ensuring that observed differences are not mistaken for inherent data quality disparities without considering distributional context.

Overall, no extreme systemic biases are immediately evident in the current results. Larger comparator groups do not consistently receive fewer flags, indicating that the severity scoring mechanism is not inherently skewed toward dominant categories. Nonetheless, the framework should be regarded as a transparent, adaptive first iteration – one that will require continued refinement with sector experts and responsiveness to the socio-political and environmental diversity of the global water sector.

> **Summary:** The framework's reliance on self-reported data without definitive benchmarks risks misclassification and demands expert oversight to guard against bias or strategic reporting. Limited gains from imputation highlight the need for adaptive, context-aware designs that balance integrity with inclusivity. Threshold sensitivity and static comparators can obscure real performance drivers, making continual refinement essential for fair, sector-responsive benchmarking.

9.2 Framing Sensitivities

Framing choices in the communication of automated anomaly detection results are not merely semantic – they shape interpretation, stakeholder trust, and the ethical legitimacy of the system. In expert feedback, terminology such as *"mild based on deviation"* was flagged as inaccessible to non-technical audiences, with suggestions for simpler, more intuitive phrasing. Similarly, explanations of imputation processes were sometimes misunderstood, leading to calls for clearer, plain-language descriptions. The challenge extends to communication with utilities: returning questionnaires with precise revision points was viewed as transparent but resource-intensive, whereas aggregated notifications risked obscuring actionable detail. This trade-off between clarity and operational feasibility shows the importance of linguistic precision. Literature on the framing and language of ethics (McNealy, 2021; Borghouts et al., 2024) cautions that wording can carry implicit value judgments, potentially biasing perception before substantive review. This dynamic is evident early in the thesis process when the term *"missing"* was debated for its perceived accusatory tone, despite its statistical appropriateness; likewise, distinctions between *"error"* and *"flag"* influence whether anomalies are interpreted as faults or as neutral signals for review.

The iterative prototyping cycle described in Chapter 4 was therefore critical in embedding ethical reflection into the design process from the outset, ensuring that transparent and context-sensitive framing evolved alongside the technical model rather than being introduced as a last-minute addition. A deliberate ethical choice during this process was to anchor explanations and flagging criteria in mathematically grounded, data-oriented terminology, thereby avoiding potentially biased or politically sensitive language. While this helped maintain neutrality, it likely contributed to expert feedback noting the current complexity of the system's framing.

In a global benchmarking context, where utilities operate under vastly different capacities, cultural norms, and linguistic interpretations, such framing sensitivities are central to ensuring that the system remains transparent, culturally aware, and ethically responsible in both its internal and external communications.

> **Summary:** The expert survey displays how anomaly detection results are framed shapes not just how they are read, but the trust and ethical legitimacy they command. In global benchmarking, where capacities, cultures, and interpretations vary widely, using clear, culturally aware language while preserving actionable detail is key to keeping the system both transparent and fair.

9.3 Trust in Automated Flagging

Trust in the automated flagging framework is not only a technical requirement but a prerequisite for its sustained adoption. In this case, expert survey feedback indicated exceptionally high institutional confidence: 100% of experts reported that they would use the system in practice, even while offering constructive points for refinement. This is significant from an ethical standpoint, as automation without stakeholder trust risks reversion to manual processes regardless of technical merit.

From an operational perspective, the framework reduces the estimated time required for technical review from approximately 75 hours to under 2 minutes[114] – a 99% efficiency gain. While this represents a substantial productivity improvement, it also prompts reflection on the ethical implications of replacing or reshaping human roles. Perfecting such a system could, in some contexts, risk displacing tasks that previously formed a substantial part of an expert's work portfolio.

At the same time, the current proposed pipeline for NewIBNET is a first-generation prototype and still depends on expert input for contextual interpretation, integration of sector knowledge, and refinement of thresholds and comparators. In this sense, the aim is not to replace expert judgment but to augment it, aligning with human-in-the-loop design principles that ensure critical oversight remains embedded in the process (Munro, 2021). Trust is therefore both a present asset and a future challenge: as technology evolves, new categories[115] emerge, and sectoral conditions shift, maintaining system reliability will require periodic recalibration. In this way, trust is not static but contingent on the framework's ability to remain accurate, relevant, and transparent in an evolving global water benchmarking landscape.

> **Summary:** Expert feedback showed unanimous willingness to use the automated flagging framework. This indicates that trust is both a technical and ethical prerequisite for its adoption. While the system delivers a 99% efficiency gain and augments rather than replaces expert judgment, sustaining that trust will require ongoing recalibration to keep it accurate, relevant, and transparent in a changing sector.

[114] The 2-minute processing time excludes communication with utilities or additional manual checks and reflects only the code runtime.

[115] The **World Bank Country and Lending Groups** classification is updated annually on July 1 (World Bank, 2024); this thesis uses the 2024–2025 version, which must likewise be updated in the system each year.

10 Conclusion

This chapter synthesises the thesis findings, drawing together technical, institutional, and ethical insights. It begins by addressing each sub-question in Section 10.1. This is followed by a critical appraisal of the study's boundaries and constraints in Section 10.2, before outlining concrete avenues for methodological, contextual, and operational enhancement in Section 10.3. The chapter concludes with final reflections in Section 10.4 that place the framework's core contributions in a global context and directly address the central research question.

10.1 Answers to Research Questions

The following section addresses each research question in turn, presenting key findings and their implications for the design, performance, and ethical positioning of the proposed anomaly detection framework.

RQ1 *How can statistical profiling and rule-based logical checks be used to detect data quality issues and prepare water utility indicator data for reliable anomaly detection?*

Chapter 5 demonstrates that statistical profiling and rule-based logical checks form the essential foundation for reliable anomaly detection, not merely as preliminary cleaning steps but as an active layer of quality assurance that forms every subsequent stage of analysis. Through systematic structural assessment, descriptive profiling, and visualisation, the statistical pipeline establishes a baseline understanding of what constitutes plausible behaviour within the NewIBNET dataset. Complementing this, the integration of domain-informed logical rules operationalises sector knowledge into reproducible checks, enabling the detection of deeply embedded inconsistencies such as implausible ratios, percentage overflows, and unit misalignments that would otherwise bypass surface-level validation. The exploration of targeted imputation strategies further indicates that completeness cannot be pursued at the expense of contextual realism, with k-NN methods proving particularly valuable in preserving variability and structural integrity in incomplete wastewater data. In practice, this layered approach strengthens the foundation for downstream anomaly detection by reducing the risk of missing, inconsistent, or implausible values entering comparative analyses.

RQ2 *How can utility metadata and comparator-based modelling support the detection of deviations in performance indicators across diverse water utilities?*

Incorporating utility metadata into comparator-based modelling proves informative in opening potential avenues for improving fairness in assessment. Peer groupings defined by operational scale and structural context allows performance to be assessed against norms that more accurately reflect a utility's real-world constraints and opportunities. Applying z-score profiling within these comparators standardises deviation measurement across diverse groups, while also revealing the method's sensitivity to imbalanced or internally heterogeneous distributions. The addition of regional and income-based comparators surfaces deviations that global averages may obscure, yet also highlights the dangers of drawing strong conclusions from sparsely populated categories. Such sparsity and imbalance can introduce significant noise and weaken statistical robustness. Comparator design emerges not as a static design choice but as a dynamic calibration exercise – one that must balance contextual richness with statistical robustness to ensure that flagged deviations remain both meaningful and actionable in diverse water utility environments.

RQ3 *Which severity scoring methodologies can be investigated to best translate statistical deviations into a prioritisation of anomalies?*

The investigation into severity scoring methodologies demonstrates that translating statistical deviations into actionable priorities requires more than fixed thresholds; it benefits from weighting schemes that reflect the reliability of the underlying comparators. Testing both simple variance-based and advanced indicator-specific covariance weightings shows that tailoring comparator influence to statistical stability and contextual relevance can reduce noise and sharpen focus on the most credible anomalies. Integrating advanced weights into a composite z-score framework condenses multiple comparator results into a single, interpretable severity metric, enabling a tiered classification from *Mild* to *Severe*. This not only streamlines reviewer workload but also preserves nuance in the assessment of anomalies, ensuring that high-severity flags represent both substantial deviation and robust supporting evidence. In practice, such a framework offers a scalable, transparent, and statistically defensible means of prioritising anomalies in diverse utility datasets, while maintaining flexibility to adjust thresholds in line with institutional risk tolerance and data quality objectives.

RQ4 *To what extent does the proposed anomaly flagging system perform reliably, and align with expert validation and benchmarking expectations?*

Evaluation of the proposed anomaly flagging system indicates that it performs reliably in detecting structural errors and logical inconsistencies, while offering the transparency and traceability needed for expert review. Internal testing shows that the multi-stage pipeline maintains consistent detection patterns across different datasets and contexts, with sensitivity to both missing and implausible values. Expert validation further confirms the system's methodological soundness and practical relevance, with unanimous agreement on its usefulness for real-world application. Feedback also highlights opportunities for refinement – particularly in clarifying the treatment of wastewater indicators, enhancing the intuitiveness of severity labels, and exploring extensions beyond anomaly detection into benchmarking and performance monitoring. Together, these findings suggest that the system aligns closely with technical, institutional, and user expectations, while retaining the adaptability needed to evolve alongside sectoral and organisational priorities.

RQ5 *What ethical and institutional implications arise from implementing an automated anomaly detection framework in the context of global water utility benchmarking?*

The ethical and institutional analysis indicates that implementing an automated framework in global water utility benchmarking is as much a governance challenge as a technical one. Reliance on self-reported data without definitive benchmarks introduces risks of misclassification, strategic reporting, and the embedding of unrepresentative baselines, making expert oversight indispensable. Design choices around imputation, comparator definitions, and severity thresholds directly shape which utilities are scrutinised, with implications for inclusivity, fairness, and political sensitivity. Framing emerges as a critical determinant of stakeholder trust, requiring language that is transparent and culturally aware while avoiding terminology that could bias interpretation. Expert feedback reveals unanimous willingness to use the system, recognising its potential to deliver substantial efficiency gains while preserving the role of human judgment in contextual interpretation. How this point is received will also depend on the audience: for managers, the main value lies in efficiency gains and resource optimisation, while for technical specialists it may raise concerns about losing human oversight, over-reliance on algorithms, or the risk of overlooking aspects that only experts can provide.

Sustaining this trust therefore requires striking a balance: highlighting the efficiency benefits while ensuring the framework remains adaptable by recalibrating thresholds, refining comparators, and evolving communication strategies in step with sectoral realities. In this way, anomaly detection can serve as a tool for equitable benchmarking rather than a source of unintended bias or exclusion.

10.2 Limitations of Current Work

Several factors constrain the scope and generalisability of the findings. These can be broadly grouped into three areas: data-specific constraints, evaluation scope, and methodological boundaries.

The anomaly detection pipeline relies exclusively on self-reported utility data, without independently verified ground truth or universally accepted benchmarks for many indicators. This creates a degree of circularity, as comparator distributions and thresholds are shaped by the quality of the submitted data – a dependency that proved stronger than initially expected. In addition, incomplete wastewater reporting prompted the use of k-NN and median imputation, which preserves partially filled submissions but leaves 127[116] utilities with fully missing values untouched. While imputation appeared promising, its overall impact on reducing participation bias is modest, showing the limits of this approach.

In the absence of a quantitative baseline, testing could not be compared against established results but instead represents a first attempt at systematic discovery. Empirical testing is limited to the

[116] Chapter 5.3 provides a detailed overview of the categorisation.

2022–2024 NewIBNET dataset and a 2024 national dataset from PERPAMSI. While both provide valuable insights, they offer only partial evidence of the framework's adaptability to other geographies, time periods, or reporting conditions. Expert validation further confirms the framework's interpretability and practical relevance, but the respondent pool is small and domain-specialised. Broader perspectives from utilities, policymakers, and actors in data-scarce environments remain underexplored.

The process is necessarily bounded by the scope and resources of a Master's thesis. Iterative refinements are made, but some design choices are validated through small expert loops rather than sector-wide trials. Likewise, while the literature review engages directly with benchmarking research, a broader comparative analysis across analogous domains could have provided additional conceptual and technical insights.

10.3 Future Work

This section outlines avenues for advancing the framework to enhance both technical robustness and ethical grounding.

Several technical refinements[117] could strengthen the framework's internal logic and adaptability. Placeholder detection warrants deeper investigation: reliance on dataset-specific proxies, such as '1.0' for unknown values, risks misclassifying legitimate entries and may be unsuitable in other benchmarking contexts. Aligning detection with international survey standards – for example, the European Social Survey's coded non-response conventions[118] – could make this process more universally applicable. Tailoring k in k-NN imputation per indicator could improve accuracy in handling missing values, while expanding beyond the z-score to alternative deviation metrics, such as Chi-square tests or non-parametric measures (Bhuyan et al., 2013), may increase robustness in sparse or skewed datasets. Comparator logic could also be made dynamic by incorporating macro-level indicators, thereby reducing biases from static grouping structures. Further, merging datasets – for example, integrating the PERPAMSI dataset with NewIBNET – would enable comparative testing across different reporting contexts. More granular bias diagnostics could identify whether specific comparator groups, regions, or income levels are disproportionately flagged, prompting targeted adjustments.

Empirical testing could be extended to a wider range of water utility datasets to evaluate the framework's adaptability and re-

[117] A detailed overview of potential framework extensions is presented in Appendix B.4.

[118] The **European Social Survey** (**ESS**) is a cross-national academic survey conducted biennially across Europe, which sets widely used standards for handling missing data. It codes specific values such as '99' or '999' to indicate non-response or missingness, offering a consistent framework that could guide placeholder detection. Read more: `https://stessrelpubprodwe.blob.core.windows.net/data/round10/survey/ESS10_data_protocol_e01_7.pdf`

silience in varied operational contexts. Such cross-utility benchmarking within the water sector would not only validate technical performance but also generate comparative insights into governance structures, framing sensitivities, and stakeholder trust. Embedding these tests in a participatory design process – engaging utility staff, policymakers, and data-scarce regions – would help ensure the framework evolves as both a technically robust and ethically grounded tool.

Beyond the validation framework itself, improvements could be made upstream[119] in the data collection process. Introducing survey-level constraints or thresholds at the point of submission could reduce the volume of implausible values reaching the flagging stage.

Ultimately, combining upstream data safeguards with targeted pipeline refinements and broader empirical validation will be essential to ensure the suggested framework remains accurate, equitable, and trusted across the diverse realities of the global water sector.

10.4 Final Remarks

> *How can data-driven mathematical models enhance validation and benchmarking of water utility indicators while ensuring reliability, decision-making integrity, and ethical transparency?*

This thesis has demonstrated that data-driven mathematical models can meaningfully enhance the validation and benchmarking of water utility indicators by combining statistical rigour with institutional awareness and ethical safeguards. The multi-stage framework developed here advances beyond conventional anomaly detection by layering structural validation (**RQ1**) with rule-based logical checks to ensure internal coherence before data reaches comparative analysis. Comparator-based modelling (**RQ2**) then situates utilities within relevant peer groups, grounding detection in contextual fairness, while severity scoring (**RQ3**) translates statistical deviations into prioritised, interpretable signals. Expert evaluation (**RQ4**) confirmed that these design choices fostered trust and usability, reinforcing the framework's alignment with principles of decision-making integrity and ethical transparency (**RQ5**).

Beyond its technical contributions, this work reveals a more am-

[119] **Out-of-Scope Extensions:** Appendix B.5 provides an overview of other extensions and observations identified during the process that lay beyond the scope of this thesis but could be explored in future iterations.

bitious vision for anomaly detection in the water industry. By now strengthening the reliability of upstream data, the framework enables more credible performance comparisons, fairer allocation of resources, and earlier identification of systemic risks. It offers not just a mechanism for flagging anomalies but a foundation for targeted capacity building, policy reform, and strategic investment – turning detection into opportunity. In contexts where access to safe water is both a human right and a driver of sustainable development, the ability to discern genuine performance gaps from data artefacts is essential. The framework's principles and methods extend beyond water utilities, offering a transferable approach for other infrastructure sectors where data quality, trust, and equity are intertwined. In this sense, the work contributes not only to refining the mechanics of benchmarking, but to shaping it as a catalyst for informed, just, and forward-looking action – moving the conversation *from red flags to real solutions*.

References

[1] Abián, David, Jorge Bernad, and Raquel Trillo-Lado (Apr. 2019). “Using contemporary constraints to ensure data consistency”. In: *SAC ’19: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pp. 2303–2310. URL: https://dl.acm.org/doi/10.1145/3297280.3297509.

[2] Ahmed, Mohiuddin, Abdun Naser Mahmood, and Jiankun Hu (Jan. 2016). “A survey of network anomaly detection techniques”. In: *Journal of Network and Computer Applications* 60, pp. 19–31. URL: https://www.sciencedirect.com/science/article/pii/S1084804515002891.

[3] Andres, Luis A. and Aroha Bahuguna (Dec. 2020). “Overcoming missing data bias in water utility indicators by using nested balanced panels”. In: *Utilities Policy* 67. URL: https://www.sciencedirect.com/science/article/pii/S095717872030103X.

[4] Andrés, Luis et al. (Oct. 2020). “Estimating the Magnitude of Water Supply and Sanitation Subsidies”. In: *World Bank*. URL: https://openknowledge.worldbank.org/server/api/core/bitstreams/0930bcfd-5843-57a4-be86-3331e9f994cc/content.

[5] Anifa, Mansurali et al. (May 2024). “Fuzzy Logic Decision Making Approach to identify the maximum influencing factor on productivity”. In: *ICIMMI ’23: Proceedings of the 5th International Conference on Information Management Machine Intelligence*, pp. 1–5. URL: https://dl.acm.org/doi/10.1145/3647444.3647837.

[6] Aubry, Mathieu (Oct. 2021). “Deep Learning for Historical Data Analysis”. In: *SUMAC’21: Proceedings of the 3rd Workshop on Structuring and Understanding of Multimedia heritAge Contents*. URL: https://dl.acm.org/doi/10.1145/3475720.3476877.

[7] Austin, Victoria Hodge Jim (Oct. 2004). “A Survey of Outlier Detection Methodologies”. In: *Artificial Intelligence Review*, pp. 85–126. URL: https://link.springer.com/article/10.1023/B:AIRE.0000045502.10941.a9.

[8] Bachinger, Florian et al. (June 2024). “Data Validation Utilizing Expert Knowledge and Shape Constraints”. In: *ACM Journal of Data and Information Quality* 16, pp. 1–27. URL: https://dl.acm.org/doi/10.1145/3661826.

[9] Batini, Carlo and Monica Scannapieca (2020). “Data Quality”. In: URL: https://link.springer.com/book/10.1007/3-540-33173-5.

[10] Baxter, Gordon and Ian Sommerville (June 2011). “Socio-technical systems: From design methods to systems engineering”. In: *Interacting with Computers* 23 (1), pp. 4–17. URL: https://ieeexplore.ieee.org/document/8147295.

[11] Belgacem, Hichem et al. (Oct. 2024). "Automated anomaly detection for categorical data by repurposing a form filling recommender system". In: *ACM Journal of Data and Information Quality* 16 (3), pp. 1–28. URL: https://dl.acm.org/doi/10.1145/3696110.

[12] Benford, Frank (Mar. 1938). "The Law of Anomalous Numbers". In: *Proceedings of the American Philosophical Society*. URL: https://isidore.co/misc/Physics%20papers%20and%20books/Zotero/storage/ZEBWDL73/Benford%20-%201938%20-%20The%20Law%20of%20Anomalous%20Numbers.pdf.

[13] Benson, Lawrence and Carsten Binnig (June 2024). "Surprise Benchmarking: The Why, What, and How". In: *DBTest '24: Proceedings of the Tenth International Workshop on Testing Database Systems*, pp. 1–8. URL: https://dl.acm.org/doi/10.1145/3662165.3662763.

[14] Berg, Caroline van den and Alexander Daninelo (Mar. 2017). "Performance of Water Utilities in Africa". In: *World Bank*. URL: https://openknowledge.worldbank.org/entities/publication/743b2631-aeb9-5998-aa7f-ebec7b9c4fab.

[15] Berg, Sanford (Apr. 2010). "Water Utility Benchmarking Measurement, Methodologies, and Performance Incentives". In: URL: https://iwaponline.com/ebooks/book/519/Water-Utility-Benchmarking-Measurement.

[16] Bhatt, Jigar D. (May 2024). "The Politics of Performance Benchmarking in Urban Water Supply: Sacrificing Equity on the Altar of Efficiency". In: *Water Alternatives* 17 (2), pp. 415–436. URL: https://www.water-alternatives.org/index.php/alldoc/articles/vol17/v17issue2/748-a17-2-6/file.

[17] Bhuyan, Monowar H., D.K. Bhattacharyya, and J.K. Kalita (June 2013). "Network Anomaly Detection: Methods, Systems and Tools". In: *IEEE Communications Surveys Tutorials* 16, pp. 303–336. URL: https://ieeexplore.ieee.org/abstract/document/6524462.

[18] Bicego, Manuele and Ferdinando Cicalese (June 2024). "Computing Random Forest-distances in the presence of missing data". In: *ACM Transactions on Knowledge Discovery from Data, Volume 18, Issue 7*, pp. 1–18. URL: https://dl.acm.org/doi/10.1145/3656345.

[19] Bingöl, Ezgi and Gizem Dürgeci Bekar (Nov. 2024). "Bi-clustering Anomaly Detection: A Dual Stage Clustering Approach Using Bayesian Gaussian Mixture Models (Bi-BGMM)". In: *2024 8th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. URL: https://ieeexplore.ieee.org/document/10757246.

[20] Borghouts, Judith et al. (Sept. 2024). "Wording Matters: The Effect of Linguistic Characteristics and Political Ideology on Resharing of COVID-19 Vaccine Tweets". In: *ACM Transactions on Computer-Human Interaction* 31 (4), pp. 1–23. URL: https://dl.acm.org/doi/10.1145/3637876.

[21] Boyatzis, Richard E. (1998). "Transforming Qualitative Information - Thematic Analysis and Code Development". In: URL: https://books.google.nl/books?id=_rfClWRhIKAC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false.

[22] Burdescu, Ruxandra et al. (Jan. 2020). "A Benchmark for the Performance of State-Owned Water Utilities in the Caribbean". In: *World Bank*. URL: https://documents1.worldbank.org/curated/en/689561580362950917/pdf/A-Benchmark-for-the-Performance-of-State-Owned-Water-Utilities-in-the-Caribbean.pdf.

[23] Buszydlik, A.J. (2024). "Finding Recourse for Algorithmic Recourse". In: URL: https://resolver.tudelft.nl/uuid:be47ad5a-5a4b-457c-b214-35c6c78cae36.

[24] Candelieri, Antonio (Mar. 2017). "Clustering and Support Vector Regression for Water Demand Forecasting and Anomaly Detection". In: *Water*. URL: https://www.mdpi.com/2073-4441/9/3/224.

[25] Castillo, Mary Geraldine and Mary Jane Samonte (Aug. 2024). "Application of Binary Classification Modelling Techniques for Water Potability Prediction". In: *ICCTA '24: Proceedings of the 2024 10th International Conference on Computer Technology Applications*, pp. 46–56. URL: https://dl.acm.org/doi/10.1145/3674558.3674566.

[26] Caylor, Justine, Somiya Metu, and Adrienne Raglin (Dec. 2020). "The Role of Multi-Criteria Decision Making in a Sentry Agents Framework Utilizing Uncertainty of Information". In: *2020 IEEE International Conference on Big Data (Big Data)*. URL: https://ieeexplore.ieee.org/document/9378243.

[27] Chandola, Varun, Arindam Banerjee, and Vipin Kumar (July 2009). "Anomaly Detection: A survey". In: *ACM Computing Surveys (CSUR)* 41, pp. 1–58. URL: https://dl.acm.org/doi/10.1145/1541880.1541882.

[28] Chang, Jingkai (Dec. 2024). "Optimization of Hadoop-Based Anomaly Detection Algorithm in Big Data Environment". In: *ICCSIE '24: Proceedings of the 2024 9th International Conference on Cyber Security and Information Engineering*, pp. 203–209. URL: https://dl.acm.org/doi/10.1145/3689236.3691499.

[29] Chitty-Vankata, Krishna Teja et al. (Feb. 2025). "LLM-Inference-Bench: Inference Benchmarking of Large Language Models on AI Accelerators". In: *SC-W '24: Proceedings of the SC '24 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis*, pp. 1362–1379. URL: https://dl.acm.org/doi/10.1109/SCW63240.2024.00178.

[30] Clifton, Chris et al. (Apr. 2022). "Differentially Private k-Nearest Neighbor Missing Data Imputation". In: *ACM Transactions on Privacy and Security* 25 (3), pp. 1–23. URL: https://dl.acm.org/doi/10.1145/3507952.

[31] Cui, Leyi (Oct. 2024). "A Formal Approach to the Analysis of Human-Machine Interaction with Fuzzy Logic". In: *SPLASH 2024: Companion Proceedings of the 2024 ACM SIGPLAN International Conference on Systems, Programming, Languages, and Applications: Software for Humanity*, pp. 31–33. URL: https://dl.acm.org/doi/10.1145/3689491.3689969.

[32] Dai, Yun and Jinghao Huang (Apr. 2021). "A Sales Forecast Method for Products with No Historical Data". In: *2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*. URL: https://ieeexplore.ieee.org/document/9442603.

[33] Deng, Wei et al. (Sept. 2022). "A consistency index analysis and calculation method for multi-measurement data". In: *2022 China International Conference on Electricity Distribution (CICED)*. URL: https://ieeexplore.ieee.org/document/9929232.

[34] Detroz, Juliana Patrícia and André Tavares da Silva (Apr. 2017). "Fraud detection in water meters using pattern recognition techniques". In: *SAC '17: Proceedings of the Symposium on Applied Computing*, pp. 77–82. URL: https://dl.acm.org/doi/10.1145/3019612.3019634.

[35] Dogo, Eustace M. et al. (June 2019). “A survey of machine learning methods applied to anomaly detection on drinking-water quality data”. In: *Urban Water Journal* 16, pp. 235–248. URL: https://www.tandfonline.com/doi/full/10.1080/1573062X.2019.1637002.

[36] Dong, Jie et al. (Aug. 2024). “Data-driven water quality prediction in Dagu River Basin, Jiaozhou Bay”. In: *ICSCIS '24: Proceedings of the 2024 International Conference on Smart City and Information System*, pp. 333–339. URL: https://dl.acm.org/doi/10.1145/3685088.3685146.

[37] Drávai, Balázs and István Z. Reguly (Feb. 2025). “Benchmarking the Evolution of Performance and Energy Efficiency Across Recent Generations of Intel Xeon Processors”. In: *SC-W '24: Proceedings of the SC '24 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis*, pp. 1413–1419. URL: https://dl.acm.org/doi/10.1109/SCW63240.2024.00182.

[38] Dudani, Jayati, Vallary Gupta, and Chirag Deb (Dec. 2022). “Building energy benchmarking in India: a discussion”. In: *BuildSys '22: Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pp. 433–438. URL: https://dl.acm.org/doi/10.1145/3563357.3566143.

[39] EPA (June 2014). “Guide for Estimating Infiltration and Inflow”. In: URL: https://www3.epa.gov/region1/sso/pdfs/Guide4EstimatingInfiltrationInflow.pdf.

[40] Foster, Vivien and Cecilia Briceño-Garmendia (2010). “Africa's Infrastructure: A Time for Transformation.” In: URL: https://documents1.worldbank.org/curated/en/246961468003355256/pdf/521020PUB0EPI11010fficial0Use0Only1.pdf.

[41] Ganjidoost, Amin and M.A. Knight and (Jan. 2018). “Benchmark Performance Indicators for Utility Water and Wastewater Pipelines Infrastructure”. In: *Journal of Water Resources Planning and Management*. URL: https://www.researchgate.net/publication/322295181_Benchmark_Performance_Indicators_for_Utility_Water_and_Wastewater_Pipelines_Infrastructure.

[42] Glickman, Edward A. (2014). “Housing Finance”. In: *An Introduction to Real Estate Finance*, pp. 335–360. URL: https://www.sciencedirect.com/science/article/pii/B9780123786265000127.

[43] Grant, Maria J. and Andrew Booth (2009). “A typology of reviews: an analysis of 14 review types and associated methodologies”. In: *Health Information Libraries Journal 26.2*, pp. 91–108. DOI: https://doi.org/10.1111/j.1471-1842.2009.00848.x.

[44] Hajirahimi, Zahra and Mehdi Khashei (July 2023). “Weighting Approaches in Data Mining and Knowledge Discovery: A Review”. In: *Neural Processing Letters*, pp. 1–46. URL: https://www.researchgate.net/publication/371982985_Weighting_Approaches_in_Data_Mining_and_Knowledge_Discovery_A_Review.

[45] Hall, Mark (Mar. 2007). “A Decision Tree-Based Attribute Weighting Filter for Naive Bayes”. In: *Knowledge-Based Systems* 20 (2), pp. 120–126. URL: https://www.sciencedirect.com/science/article/pii/S0950705106002000.

[46] Herrera-Viedma, Enrique et al. (Dec. 2020). “Revisiting Fuzzy and Linguistic Decision Making: Scenarios and Challenges for Making Wiser Decisions in a Better Way”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 51 (1), pp. 191–208. URL: https://ieeexplore.ieee.org/document/9306916.

[47] Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: URL: https://www.researchgate.net/publication/13853244_Long_Short-Term_Memory.

[48] Huang, Mi, LingLi Li, and Ping Xuan (Mar. 2019). "Evaluating Data Consistency with Matching Dependencies from Multiple Sources". In: *2019 IEEE International Conference on Power Data Science (ICPDS)*. URL: https://ieeexplore.ieee.org/document/9017191.

[49] Jalal, Dziri and Tahar Ezzedine (June 2020). "Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks". In: *2020 International Wireless Communications and Mobile Computing (IWCMC)*. URL: https://ieeexplore.ieee.org/document/9148431.

[50] Jia, Xudong (Mar. 2022). "Detecting Water Quality Using KNN, Bayesian and Decision Tree". In: *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*. URL: https://ieeexplore.ieee.org/document/9852554.

[51] Jiang, Minqi, Songqiao Han, and Hailiang Huang (Aug. 2023). "Anomaly Detection with Score Distribution Discrimination". In: *KDD '23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 984–996. URL: https://dl.acm.org/doi/10.1145/3580305.3599258.

[52] Jones, Eric (Aug. 2023). "Data Accuracy vs Data Integrity: Similarities and differences". In: URL: https://www.ibm.com/think/topics/data-accuracy-vs-data-integrity.

[53] Kahla, Mayssa Ben et al. (Jan. 2025). "Enhanced Fuzzy Score-Based Decision Support System for Early Stroke Prediction". In: *ACM Transactions on Computing for Healthcare* 6 (1), pp. 1–23. URL: https://dl.acm.org/doi/10.1145/3703461.

[54] Kang, Gaganjot, Jerry Zeyu Gao, and Gang Xie (June 2017). "Data-driven Water Quality Analysis and Prediction: A Survey". In: URL: https://ieeexplore.ieee.org/document/7944943.

[55] Kanyama, M.N. et al. (June 2024). "Machine Learning Applications for Anomaly Detection in Smart Water Metering Networks: A Systematic Review". In: *Physics and Chemistry of the Earth, Parts A/B/C* 134. URL: https://www.sciencedirect.com/science/article/pii/S1474706524000160.

[56] Kim, Harim, Chang Ha Lee, and Charmgil Hong (Oct. 2024). "Transformer for Point Anomaly Detection". In: *CIKM '24: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1080–1088. URL: https://dl.acm.org/doi/10.1145/3627673.3679859.

[57] Kumar, Sushant et al. (May 2024). "Applications, Challenges, and Future Directions of Human-in-the-Loop Learning". In: *IEEE Access* 12, pp. 75735–75760. URL: https://ieeexplore.ieee.org/document/10530996.

[58] Kutz, Jim (Aug. 2025). "Data Validation in ETL: Why It Matters and How to Do It Right". In: URL: https://airbyte.com/data-engineering-resources/data-validation.

[59] Li, Hao et al. (Dec. 2023). "Application of Comprehensive Water Quality Labeling Index Method in Water Quality Evaluation of Xiangjiang River". In: *2023 4th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+AI)*. URL: https://ieeexplore.ieee.org/document/10551121.

[60] Lim, Haksoo et al. (Oct. 2023). "MadSGM: Multivariate Anomaly Detection with Score-based Generative Models". In: *CIKM '23: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 1411–1420. URL: https://dl.acm.org/doi/10.1145/3583780.3614956.

[61] Lin, Guo and Yongfeng Zhang (Feb. 2025). "Fuzzy Neural Logic Reasoning for Robust Classification". In: *ACM Transactions on Knowledge Discovery from Data* 19 (2), pp. 1–29. URL: https://dl.acm.org/doi/10.1145/3704728.

[62] Liu, Changjie and Xianning Lin (Jan. 2025). "Multi-classifier semi-supervised data stream classification algorithm based on online learning". In: *CECCT '24: Proceedings of the 2024 2nd International Conference on Electronics, Computers and Communication Technology*, pp. 127–132. URL: https://dl.acm.org/doi/10.1145/3705754.3705777.

[63] Liu, Nan and Christopher C. Yang (May 2007). "A link classification based approach to website topic hierarchy generation". In: *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pp. 1127–1128. URL: https://dl.acm.org/doi/10.1145/1242572.1242728.

[64] Liu, Yaqi, Xinyi Liu, and Shujie Wang (Mar. 2025). "Research and Implementation of the front-end and back-end data validation method based on metadata". In: *ICAICE '24: Proceedings of the 5th International Conference on Artificial Intelligence and Computer Engineering*, pp. 593–597. URL: https://dl.acm.org/doi/10.1145/3716895.3717000.

[65] Manghee, Seema and Alice Poole (Aug. 2012). "Approaches to Conducting Political Economy Analysis in the Urban Water Sector". In: *World Bank*. URL: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/560131468339257950/approaches-to-conducting-political-economy-analysis-in-the-urban-water-sector.

[66] Mauro, Francesco and Benjamin Rich (July 2023). "SEN2DWATER: A Novel Multispectral and Multitemporal Dataset and Deep Learning Benchmark for Water Resources Analysis". In: *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*. URL: https://ieeexplore.ieee.org/document/10282352.

[67] McCabe, T.J. (Dec. 1976). "A Complexity Measure". In: *IEEE Transactions on Software Engineering* SE-2, pp. 308–320. URL: https://ieeexplore.ieee.org/document/1702388.

[68] McNealy, Jasmine E. (Sept. 2021). "Framing and Language of Ethics: Technology, Persuasion, and Cultural Context". In: *Journal of Social Computing* 2 (3), pp. 226–237. URL: https://ieeexplore.ieee.org/document/9684740.

[69] Meng, Mei et al. (July 2021). "Fuzzy Comprehensive Evaluation of Performance on Urban Water Supply System". In: *2021 40th Chinese Control Conference (CCC)*. URL: https://ieeexplore.ieee.org/document/9550032.

[70] Miao, Xiaoye et al. (June 2024). "An Experimental Survey of Missing Data Imputation Algorithms". In: *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. URL: https://ieeexplore.ieee.org/document/10597832.

[71] Misra, Alka et al. (Dec. 2024). "Benchmarking of Government Websites: An approach towards good governance". In: *ICEGOV '24: Proceedings of the 17th International Conference on Theory and Practice of Electronic Governance*, pp. 184–189. URL: https://dl.acm.org/doi/10.1145/3680127.3680150.

[72] Mitchell, Tom (1997). “Machine Learning”. In: URL: https://www.cs.cmu.edu/~tom/mlbook.html.

[73] Montgomery, Douglas C and George C Runger (Jan. 2014). “Applied Statistics and Probability for Engineers”. In: URL: https://kolegite.com/EE_library/books_and_lectures/%D0%9C%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0/Applied%20Statistics%20and%20Probability%20for%20Engineering.pdf.

[74] Moses, Barr (July 2025). “What is Data Reliability and How do I Make my Data Reliable?” In: URL: https://www.montecarlodata.com/blog-what-is-data-reliability.

[75] Munro, Rob (2021). “Human-in-the-loop Machine Learning: Active Learning and Annotation for Human-Centred AI”. In: URL: https://ieeexplore.ieee.org/document/10280384.

[76] Murtagh, Jack (May 2023). “What Is Benford’s Law? Why This Unexpected Pattern of Numbers Is Everywhere”. In: URL: https://www.scientificamerican.com/article/what-is-benfords-law-why-this-unexpected-pattern-of-numbers-is-everywhere.

[77] Nofal, Samer, Abdullah Alfarrarjeh, and Amani Abu Jabal (July 2021). “A use case of anomaly detection for identifying unusual water consumption in Jordan”. In: *Water Supply* 22, pp. 1131–1140. URL: https://iwaponline.com/ws/article/22/1/1131/82887/A-use-case-of-anomaly-detection-for-identifying.

[78] North Carolina Water Service (2024). “How Utility Systems Work”. In: URL: https://www.myutility.us/carolinawater/water-smart/utility-systems.

[79] Pang, Guansong et al. (Mar. 2021). “Deep Learning for Anomaly Detection: A Review”. In: *ACM Computing Surveys (CSUR)* 54, pp. 1–38. URL: https://dl.acm.org/doi/abs/10.1145/3439950.

[80] Paul B. Bokingkito, Jr. and Lomesindo T. Caparida (Oct. 2018). “Using Fuzzy Logic for Real - Time Water Quality Assessment Monitoring System”. In: *ICACR 2018: Proceedings of the 2018 2nd International Conference on Automation, Control and Robots*, pp. 21–25. URL: https://dl.acm.org/doi/10.1145/3293688.3293695.

[81] Peleska, Jan et al. (June 2021). “Efficient data validation for geographical interlocking systems”. In: *Formal Aspects of Computing* 33, pp. 925–955. URL: https://dl.acm.org/doi/10.1007/s00165-021-00551-6.

[82] Perez, Juan (Apr. 2023). “How Automation Drives Business Growth and Efficiency”. In: URL: https://hbr.org/sponsored/2023/04/how-automation-drives-business-growth-and-efficiency.

[84] Qi, Daqian, Binglei Chang, and Wenwen Li (July 2024). “Big Data Analysis and Intelligent Decision Support System for Environmental Water Quality: Application of Artificial Intelligence in Water Environmental Protection”. In: *2024 3rd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS)*. URL: https://ieeexplore.ieee.org/document/10708824.

[85] Raciti, Massimiliano, Jordi Cucurull, and Simin Nadjm-Tehrani (2012). “Anomaly Detection in Water Management Systems”. In: *Critical Infrastructure Protection*, pp. 98–119. URL: https://link.springer.com/chapter/10.1007/978-3-642-28920-0_6.

[86] Radanliev, Petar (Aug. 2024). "AI Ethics: Integrating Transparency, Fairness, and Privacy inAI Development". In: *Applied Artificial Intelligence* 39. URL: https://www.tandfonline.com/doi/epdf/10.1080/08839514.2025.2463722?needAccess=true.

[87] Riley, Christine, Bo Xie, and Anjum Khurshid (Oct. 2021). "Challenges encountered in comparing international policy responses to COVID-19 and their effects". In: *Health Research Policy and Systems*. URL: https://health-policy-systems.biomedcentral.com/articles/10.1186/s12961-021-00783-1?.

[88] Russel, Stuart and Peter Norvig (2006). "Artificial Intelligence: A Modern Approach". In: URL: https://aima.cs.berkeley.edu/.

[89] Sarkar, P. (2022). "Perceptions of mathematical fairness notions by hiring professionals". In: URL: https://resolver.tudelft.nl/uuid:521e8b01-2f5a-4717-9a26-9fe2636b8427.

[90] Senanayake, S.M.C. Prageeth and M. Anusha Wijewardane (July 2012). "Benchmarking Energy and Water Consumption of Supermarkets in Sri Lanka". In: *Moratuwa Engineering Research Conference (MERCon)*. URL: https://ieeexplore.ieee.org/document/9906145.

[91] Shankar, Shreya, Labib Fawaz, and Aditya Parameswaran Karl Gyllstrom (Oct. 2023). "Automatic and Precise Data Validation for Machine Learning". In: *CIKM '23: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 2198–2207. URL: https://dl.acm.org/doi/10.1145/3583780.3614786.

[92] Sharma, Asheesh et al. (Aug. 2012). "Fuzzy logic application in water supply system management: A case study". In: *2012 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS)*. URL: https://ieeexplore.ieee.org/document/6291060.

[93] Silalahi, Swardiantara et al. (May 2024). "Severity-Oriented Multiclass Drone Flight Logs Anomaly Detection". In: *IEEE Access* 12. URL: https://ieeexplore.ieee.org/document/10520297.

[94] Silva, W.S.D. De, A. H. R. Ratnasooriya, and Harsha Abeykoon (June 2023). "Non-Revenue Water Reduction Strategies for an Urban Water Supply Scheme: A Case Study for Gampaha Water Supply Scheme". In: *2023 Moratuwa Engineering Research Conference (MERCon)*. URL: https://ieeexplore.ieee.org/document/10355422.

[95] Singhal, A. and D.E. Seborg (June 2001). "Matching patterns from historical data using PCA and distance similarity factors". In: *Proceedings of the 2001 American Control Conference. (Cat. No.01CH37148)*. URL: https://ieeexplore.ieee.org/document/945986.

[96] Sommerville, Ian and Guy Dewsbury (July 2007). "Dependable domestic systems design: A socio-technical approach". In: *Interacting with Computers* 19 (4), pp. 438–456. URL: https://ieeexplore.ieee.org/document/8147154.

[97] Song, Jie and Yeye He (June 2021). "Auto-Validate: Unsupervised Data Validation Using Data-Domain Patterns Inferred from Data Lakes". In: *SIGMOD '21: Proceedings of the 2021 International Conference on Management of Data*, pp. 1678–1691. URL: https://dl.acm.org/doi/10.1145/3448016.3457250.

[98] Taleb, Ikbal, Mohamed Adel Serhani, and Rachida Dssouli (Nov. 2018). "Big Data Quality Assessment Model for Unstructured Data". In: *2018 International Conference on Inno-*

vations in Information Technology (*IIT*). URL: https://ieeexplore.ieee.org/document/8605945.

[99] Tong, Wu et al. (Mar. 2009). "Product Innovation Design and Design Management". In: *2009 First International Workshop on Education Technology and Computer Science*. URL: https://ieeexplore.ieee.org/document/4958773.

[100] Troyanskaya, O., M. Cantor, and G. Sherlock (June 2001). "Missing value estimation methods for DNA microarrays". In: *Bioinformatics*. URL: https://pubmed.ncbi.nlm.nih.gov/11395428/.

[101] Tsagarakis, Konstantinos (Jan. 2018). "Operating Cost Coverage vs. Water utility complaints". In: 10 (1). URL: https://www.scopus.com/record/display.uri?eid=2-s2.0-85039928678&origin=resultslist&sort=plf-f&src=s&sot=b&sdt=b&s=TITLE-ABS-KEY%28ibnet%29&relpos=9.

[102] Tsai, Emmali (Jan. 2025). "Data-Driven Trust: what we can and cannot see in water data". In: URL: https://www.policyinnovation.org/insights/data-driven-trust.

[103] Tukey, John (1977). "Exploratory Data Analysis". In: URL: https://consoleflare.com/blog/wp-content/uploads/2022/09/Exploratory-Data-Analysis-1977-John-Tukey.pdf.

[104] United Nations (2025). "Sustainable Development Goals: Goal 6 - Ensure access to water and sanitation for all". In: URL: https://www.un.org/sustainabledevelopment/water-and-sanitation/.

[83] van de Poel, Ibo and Lamber Royakkers (Feb. 2011). "Ethics, Technology, and Engineering: An Introduction". In: URL: https://cdn.prexams.com/6229/BOOK.pdf.

[105] Waller, Madeleine et al. (Jan. 2025). "Bias Mitigation Methods: Applicability, Legaily, and Recommendations for Development". In: *Journal of Artificial Intelligence Research* 81. URL: https://dl.acm.org/doi/10.1613/jair.1.16759.

[106] Wang, Sisi et al. (n.d.). "Fuzzy Weighted Principal Component Analysis for Anomaly Detection". In: *ACM Transactions on Knowledge Discovery from Data* 19 (3), pp. 1–22. URL: https://dl.acm.org/doi/10.1145/3715148.

[107] Wang, Xinyue, Hafiz Asif, and Jaideep Vaidya (Apr. 2023). "Preserving Missing Data Distribution in Synthetic Data". In: *WWW '23: Proceedings of the ACM Web Conference 2023*, pp. 2110–2121. URL: https://dl.acm.org/doi/10.1145/3543507.3583297.

[108] Wibowo, Santoso and Srimannarayana Grandhi (June 2017). "Evaluating the sustainability performance of urban water services". In: *2017 12th IEEE Conference on Industrial Electronics and Applications* (*ICIEA*). URL: https://ieeexplore.ieee.org/document/8282849.

[109] Wibowo, Santoso and Srimannarayana Grandhi (June 2018). "Multicriteria assessment of urban water supply system providers". In: *2018 13th IEEE Conference on Industrial Electronics and Applications* (*ICIEA*). URL: https://ieeexplore.ieee.org/document/8397937.

[110] Winona, Avelisa Yoelma and Didit Adytia (Aug. 2020). "Short Term Forecasting of Sea Level by Using LSTM with Limited Historical Data". In: *2020 International Conference*

on Data Science and Its Applications (ICoDSA). URL: https://ieeexplore.ieee.org/document/9213025.

[111] World Bank (Apr. 2011). "World Development Report: Conflict, Security, and Development". In: URL: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/806531468161369474/world-development-report-2011-conflict-security-and-development-overview.

[112] World Bank (2014). "The IBNET Water Supply and Sanitation Blue Book 2014". In: URL: https://openknowledge.worldbank.org/server/api/core/bitstreams/acc712cd-bec2-517c-8859-557925dca966/content.

[113] World Bank (June 2022). "Improving Data Quality for an Effective Social Registry in Indonesia". In: URL: https://documents1.worldbank.org/curated/en/099515110132223290/pdf/P177341026be35061089cf0d835280eaf15.pdf.

[114] World Bank (July 2023). "Water". In: URL: https://www.worldbank.org/en/topic/water/overview.

[115] World Bank (2024). "World Bank Country and Lending Groups". In: URL: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups.

[116] World Health Organisation (Sept. 2023). "Drinking-water". In: URL: https://www.who.int/news-room/fact-sheets/detail/drinking-water.

[117] Wu, Zheng Yi, M.ASCE, and Yekun He (July 2021). "Time Series Data Decomposition-Based Anomaly Detection and Evaluation Framework for Operational Management of Smart Water Grid". In: *9* 147. URL: https://ascelibrary.org/doi/full/10.1061/%28ASCE%29WR.1943-5452.0001433.

[118] Xian, Feng and Yisu Yin (Dec. 2024). "Intelligent Decision Optimization Method for Water Information System Based on Artificial Intelligence". In: *AI2A '24: Proceedings of the 2024 4th International Conference on Artificial Intelligence, Automation and Algorithms*, pp. 117–121. URL: https://dl.acm.org/doi/10.1145/3700523.3700545.

[119] Yang, Jun-yao and Jian-fei Yue (Mar. 2011). "Comprehensive judgement for urban water supply performance based on the SE-DEA model". In: *2011 3rd International Conference on Computer Research and Development*. URL: https://ieeexplore.ieee.org/document/5763928.

[120] Yi, Wonjun, Wonho Jung, and Yong-Hwa Park (Nov. 2024). "Performance Metric for Multiple Anomaly Score Distributions with Discrete Severity Levels". In: *IECON 2024 - 50th Annual Conference of the IEEE Industrial Electronics Society*. URL: https://ieeexplore.ieee.org/document/10905500.

[121] Yu, Ziwen et al. (Aug. 2023). "Water Quality Classification Evaluation based on Water Quality Monitoring Data". In: *2023 11th International Conference on Information Systems and Computing Technology (ISCTech)*. URL: https://ieeexplore.ieee.org/document/10438258.

[122] Yumang, Analyn, Lois Anne Borrega, and Almer Dave Dizon (Oct. 2021). "Evaluation of Water Filtration System Through Water Quality Testing Using Fuzzy Logic". In: *ICIEI '21: Proceedings of the 6th International Conference on Information and Education Innovations*, pp. 104–111. URL: https://dl.acm.org/doi/10.1145/3470716.3470734.

[123] Zeng, Congyi, Yongsheng Ding, and Kuangrong Hao (Mar. 2012). “Decision Support System for Water Allocation Based on Similarity Measurement”. In: *2012 International Conference on Computer Science and Electronics Engineering*. URL: https://ieeexplore.ieee.org/document/6188159.

[124] Zhu, Changjun (July 2009). “Fuzzy Neural Network Model and Its Application in Water Quality Evaluation”. In: *2009 International Conference on Environmental Science and Information Application Technology*. URL: https://ieeexplore.ieee.org/document/5200111.

Appendices

A
Data Foundations

A.1 Input, Indicator, and Results Tables

This section provides core reference tables. It contains the indicator equations used for logic validation, a missing-data review for the 2022–2024 dataset, a descriptive statistical summary of the calculated indicators, covariance-sum normalised weights per indicator and comparator, and finally a list of all raw input questions

Table A.1: *Overview of Indicators and Equations where Qx aligns with the raw input names*

I	Name	Equation
1	Percentage of Drinking Water Coverage	$\frac{Q2}{Q1} * 100$
2	Continuity of Supply (Hours per Day)	$Q3$
3	Percentage of Customers with 24/7 Supply	$\frac{Q4}{Q2} * 100$
4	Non Revenue Water	$\frac{(Q6-Q7)*1000}{Q8*24}$
5	Percentage of Non Revenue Water	$\frac{Q6-Q7}{Q6} * 100$
6	Percentage of Sanitation Coverage	$\frac{Q10}{Q9} * 100$
7	Sewer Blockages	$\frac{Q11}{Q12} * 100$
8	Percentage of Wastewater Collected and Treated	$\frac{Q14}{Q13} * 100$
9	Percentage of Collection Rate	$\frac{Q15}{Q16} * 100$
10	Percentage of Metered Connections	$\frac{Q17}{Q8} * 100$
11	Percentage of Service Complaints Resolved	$\frac{Q19}{Q18} * 100$
12	Percentage of Drinking Water Quality	$\frac{Q21}{Q20} * 100$
13	Percentage of Operation Cost Coverage	$\frac{Q22}{Q16} * 100$
14	Number of employees per 1000 connections	$\frac{Q23}{Q8+Q24} * 1000$
15	Percentage of Female Employees	$\frac{Q25}{Q23} * 100$

Table A.2: *Missing Data Overview in Raw Data Values NewIBNET 2022–2024 (Q1–Q20)*

	Name	Number of Missing Values	Number of Placeholder Values
Q1	Total Population	0	1
Q2	Total Population Service Size	0	4
Q3	Average Daily Supply	0	8
Q4	Total Customers with 24/7 Supply	0	26
Q6	Total Water Volume	1	1
Q7	Total Billed Water Volume	0	2
Q8	Total Water Connections	14	4
Q9	Total Population for Wastewater Services	132	5
Q10	Total Service Population for Wastewater Services	131	14
Q11	Number of Sewer Blockages	138	16
Q12	Sewer Pipe Length	139	7
Q13	Total Wastewater Volume	138	12
Q14	Total Treated Wastewater Volume	136	25
Q15	Total Revenue Billed	0	6
Q16	Total Revenue Collected	0	8
Q17	Total Metered Connections	0	25
Q18	Total Complaints Received	0	10
Q19	Total Complaints Resolved	0	10
Q20	Number of Water Samples	0	9

Table A.3: *Missing Data Overview in Raw Data Values NewIBNET 2022–2024 (Q21–Q25)*

	Name	Number of Missing Values	Number of Placeholder Values
Q21	Water Samples Meeting Requirements	0	9
Q22	Operating Expenses	0	9
Q23	Number of Fulltime Employees	8	6
Q24	Total Wastewater Connections	138	13
Q25	Number of Female Employees	0	18

Table A.4: *Indicator Descriptive Statistics (Min, Max, Mean, Median)*

Indicator	Min	Max	Mean	Median
I1	0.00	100.00	69.48	78.00
I2	1.00	24.00	17.57	20.00
I3	0.00	100.00	43.89	25.50
I4	0.00	16038805.97	129782.76	3026.85
I5	0.00	99.98	38.33	35.90
I6	0.00	100.00	55.58	54.00
I7	0.26	480769.23	6946.43	225.62
I8	0.00	275.89	85.55	100.00
I9	0.00	666666.67	2583.61	95.65
I10	0.00	100.00	77.39	99.77
I11	0.00	100.00	87.21	96.91
I12	0.00	100.00	88.18	97.86
I13	0.00	58000819200.00	216572271.97	93.37
I14	0.00	1102.04	19.53	4.60
I15	0.25	100.00	24.28	21.55

Table A.5: *Covariance-Sum Normalised Weights per Indicator and Comparator*

I	Global	Population	Connections	Region	Income
I1	0.2147	0.2188	0.1829	0.1881	0.1956
I2	0.2144	0.2154	0.1911	0.1912	0.1880
I3	0.2022	0.2059	0.2119	0.1945	0.1856
I4	0.1904	0.2105	0.2042	0.1978	0.1971
I5	0.2009	0.2069	0.1988	0.1956	0.1978
I6	0.2533	0.2440	0.1782	0.1667	0.1577
I7	0.0716	0.1230	0.2406	0.2757	0.2891
I8	0.2832	0.2616	0.1510	0.1582	0.1459
I9	0.2653	0.2255	0.1167	0.2148	0.1778
I10	0.2104	0.2079	0.2009	0.1893	0.1915
I11	0.2219	0.2203	0.1887	0.1850	0.1840
I12	0.2311	0.2245	0.1928	0.1801	0.1715
I13	0.1982	0.1993	0.1905	0.1906	0.2215
I14	0.2805	0.1465	0.1757	0.1876	0.2096
I15	0.2069	0.2129	0.2124	0.1801	0.1877

Table A.6: *Part 1: Overview of the list of questions asked within the NewIBNET survey (KPI)*

Name	**Used in Indicator**	**Full Question**	**Datatype**
Name	-	The name of the utility.	String
Country	-	The country the utility is based in.	String
Region	-	The region the utility is based in.	String
Build Date	-	When was this utility built?	Date
Survey Completion	-	When was this survey completed?	Date
Structure of Utility	-	Based on a set of options, what is the structure of the utility?	Choice
Nature of Service Area	-	What is the nature of the utility's service area?	Choice
Network of Pipes	-	Does the utility supply water to a customer tap through a network of pipes?	Choice
Treating Water	-	Does the utility treat its drinking water before supply?	Choice
Wastewater Treatment Services	-	Does this utility provide wastewater treatment services?	Choice
Wastewater Collection	-	Does the utility collect wastewater?	Choice
Collect before Discharge	-	Does the utility treat collected wastewater before discharge?	Choice
Location of discharge	-	Where is the effluent discharged?	Choice
Q1 Drinking Water Coverage: Total Population in Service Area	✓	What is the total population in your water service area?	Numerical
Q2 Drinking Water Coverage Population with Water Services	✓	What is the population in the service area with water services from the utility?	Numerical
Q8 Drinking Water Coverage: Total Water Service Connections	✓	What is the total number of water service connections?	Numerical
Drinking Water Coverage: Direct Household Connections	-	What is the total number of direct household water connections?	Numerical
Drinking Water Coverage: Public Tap Water Connections	-	What is the total number of public tap/standpoint water connections?	Numerical
Drinking Water Coverage: Commerical+	-	What is the total number of commerical, institutional, industrial, and other water connections?	Numerical
Q3 Continuity of Supply: Average Daily Supply	✓	What is the average daily supply in hours per day units?	Numerical
Q4 Percentage with Supply: Number of Customers	✓	What is the total number of customers that are supplied with service 24 hours per day, seven days per week?	Numerical
Q6 Non-Revenue Water: Produced Water Volume	✓	What is the total produced water volume?	Numerical
Q7 Non-Revenue Water: Total Billed Water Volume	✓	What is the total water volume billed?	Numerical
Non-Revenue Water: Volume Residential Private	-	What is the volume of water sold to residential customers through direct connections?	Numerical
Non-Revenue Water: Volume Residential Public	-	What is the volume of water sold to residential customers through public tap/standpoint water connections?	Numerical
Non-Revenue Water: Volume Commerical+	-	What is the volume of water sold to commercial, institutional, industrial, and other water connections?	Numerical
Non-Revenue Water: Total Length Network	-	What is the total length of the water distribution network?	Numerical
Non-Revenue Water: Pipe Breaks	-	What is the total number of pipe breaks recorded for the water network?	Numerical

Table A.7: *Part 2: Overview of the list of questions asked within the NewIBNET survey (KPI)*

Name	Used in Indicator	Full Question	Datatype
Q9 Sanitation Coverage: Total Population in Wastewater Service Area	✓	What is the total population in your wastewater service area?	Numerical
Q10 Sanitation Coverage: Total Utility Population in Wastewater Service Area	✓	What is the population in the service area with wastewater services from the utility?	Numerical
Q24 Sanitation Coverage: Total Wastewater Service Connections	✓	What is the total number of wastewater service connections?	Numerical
Sanitation Coverage: Total Direct Wastewater Connections	-	What is the total number of direct household wastewater connections?	Numerical
Sanitation Coverage: Total Other Wastewater Connections	-	What is the total number of commercial, institutional, industrial, and other wastewater connections?	Numerical
Q11 Sewer Blockages: Total Blockages	✓	What was the the total number of blockages?	Numerical
Q12 Sewer Blockages: Total Length Sewer Network	✓	What is the total length of pipe in the sewer network?	Numerical
Q13 Wastewater Collected & Treated: Wastewater through Tanks and Sewers	✓	What is the volume of collected wastewater through piped sewerage system or tankers?	Numerical
Q14 Wastewater Collected & Treated: Volume Wastewater	✓	What is the volume of collected wastewater that is treated?	Numerical
Q16 Revenue Collection Rate: Total Revenue Collected	✓	What was the total revenue collected?	Numerical
Q15 Revenue Collection Rate: Total Revenue Billed	✓	What is the total revenue billed?	Numerical
Q17 Percentage of Metered Connections	✓	What are the total number of metered connections?	Numerical
Q18 Service Complaints Resolved: Total Customer Complaints	✓	What was the total number of customer complaints received?	Numerical
Q19 Service Complaints Resolved: Total	✓	What was the total number of customer complaints resolved?	Numerical
Q20 Drinking Water Quality: Water Samples	✓	How many water samples were taken?	Numerical
Q21 Drinking Water Quality: Samples Meeting Guidelines	✓	How many water samples met all required guidelines?	Numerical
Q22 Operation Cost Coverage: Total Operating Expenses	✓	What were the total operating expenses?	Numerical
Operation Cost Coverage: Labor Expenses	-	What were total labor expenses?	Numerical
Operation Cost Coverage: Energy Expenses	-	What were total energy expenses?	Numerical
Operation Cost Coverage: Other Expenses	-	What were total other expenses (production, chemicals, maintenance, administrative, etc.)?	Numerical
Q23 Number of Employees: Total Full-time Employees	✓	What was the total number of full time employees?	Numerical
Number of Employees: Total Full-time Equivalent Employees	-	What was the total number of full-time equivalent employees?	Numerical
Number of Employees: Full-time Managers	-	What was the total number of full-time managers?	Numerical

Table A.8: *Part 3: Overview of the list of questions asked within the NewIBNET survey (KPI)*

Name	Used in Indicator	Full Question	Datatype
Q25 Percentage of Female Employees: Total Full-time Women	✓	What was the total number of full-time employees that are women?	Numerical
Percentage of Female Employees: Total Full-time Equivalent Women	-	What was the total number of full-time equivalent employees that are women?	Numerical
Percentage of Female Employees: Total Full-time Female Managers	-	What was the total number of full-time managers that are women?	Numerical

B
Methodological Additions

This appendix chapter presents additional methodological work that, while outside the main scope of the thesis, is carried out to enrich the study. It includes a comparative analysis with other benchmarking platforms, further detail on computational procedures and potential technical improvements, an extended exploration of Benford's Law, and several extra and out-of-scope extensions. These additions provide complementary insights that may be of interest for future research and system development.

B.1 Comparative Analysis with Benchmarking Platforms

To contextualise the architectural and strategic decisions underpinning this prototype, a comparative review is conducted with two prominent benchmarking systems: the American Water Works Association (AWWA) Utility Benchmarking Program and the European Benchmarking Co-operation (EBC). These platforms are selected due to their maturity, thematic relevance, and availability of documentation – complemented by direct communication with AWWA. While all three initiatives – NewIBNET, AWWA, and EBC – share a common goal of enabling performance benchmarking for water and wastewater services, they diverge in structure, depth, and strategic orientation.

At a foundational level, all three frameworks rely on core operational KPIs and pursue comparative performance analysis. However, their design choices reflect distinct institutional priorities. An overview of this comparison between the three benchmarking systems is shown in Figure B.1.

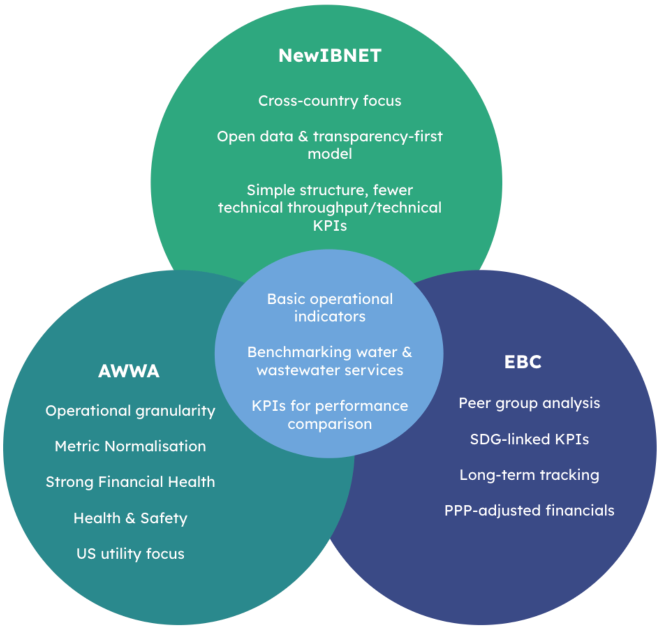


Figure B.1: *The Venn diagram illustrates the unique and shared attributes of the three benchmarking platforms: NewIBNET, EBC, and AWWA. The overlapping sections represent common features across platforms, while the non-overlapping areas highlight attributes unique to each.*

AWWA adopts an operational and performance-centric approach, geared toward technical optimisation and financial robustness within U.S.-based utilities. Its focus on metric normalisation, health and safety indicators, and optional advanced modules positions it as a tool for fine-tuned internal improvement.

EBC, by contrast, is more policy- and collaboration-driven – promoting peer-group learning, long-term trend tracking, and integration with Sustainable Development Goals (SDGs). Its strength lies in systemic benchmarking and strategic sector reform, particularly across European utilities.

In contrast, NewIBNET has historically embraced an inclusion-first model – prioritising accessibility, open data, and cross-country comparability across a highly heterogeneous set of utilities. While this approach supports data democratisation, it also constrains technical depth and comparative precision, particularly in its reliance on global averages and its limited treatment of throughput or context-sensitive indicators.

This analysis reveals concrete design innovations that NewIBNET could consider in future iterations of survey deployment. From EBC, features such as time-series analytics[120], purchasing power parity (PPP) adjustments[121], SDG-aligned indicators, and anonymous peer-group dashboards offer scalable ways to integrate strategic benchmarking without compromising usability. From AWWA, upstream survey design could be enhanced through tiered operational metrics, resilience and sustainability indicators, and clearer missing data protocols. Notably, AWWA reported that utilities are not required to complete all questions[122], and that outliers are flagged both automatically and manually – a hybrid approach that may inspire future human-in-the-loop enhancements. Furthermore, AWWA is developing an incentive mechanism for utilities surpassing 90% completion, highlighting the role of behavioural nudges in improving data coverage, a topic underexplored in NewIBNET's current model.

Ultimately, while NewIBNET operates within a more resource-constrained and globally distributed ecosystem, this comparative lens highlights not only the limitations but also the unique positioning of its benchmarking strategy. Its simplicity and open access model remain powerful, but future evolutions may benefit from selective borrowing – adopting modular enhancements that remain aligned with its core mission of equitable and transparent global benchmarking.

B.2 Computational Analysis

Linking back to Chapter 8, which discusses the computational efficiency and performance of the proposed pipeline, certain stages exhibit relatively high cyclomatic complexity. Although the system currently runs in under two minutes, it is important to consider technical improvements for future iterations, particularly as NewIBNET has the potential to scale to more utilities and larger datasets.

The most impactful optimisations for achieving a more efficient performance include:

- Pre-loading all indicator and questionnaire datasets once, avoiding repeated disk I/O.
- Vectorising **Stage 2** (**Indicator Logic**) to replace per-utility loops with column-wise operations.

[120] **Time-Series Analytics:** Involves analysing data points collected or recorded at successive time intervals to identify trends, patterns, and seasonal effects. Discussed briefly within Chapter 3 with reference to Wu et al., 2021's work.

[121] **Purchasing Power Parity (PPP) Adjustments:** Account for differences in price levels between countries, allowing for more accurate cross-country economic comparisons.

[122] **AWWA Median Completion Rate:** Reported to be 91% (2025)

- Pre-computing comparator statistics (means, standard deviations) once for **Stage 3** (**Comparator Analysis and Severity Scoring**).
- Placing z-score outputs into a single structured table to reduce file overhead.

These measures can reduce runtime by factors of 3–10 and lower complexity, improving maintainability and reducing implementation risks. In effect, runtime profiling identifies *where* to optimise (**Stage 3**), while complexity analysis clarifies *what* to refactor (**Stages 2–3**), ensuring the pipeline remains performant and sustainable as operations scale.

B.3 Benford's Law

Benford's Law, which describes the expected frequency distribution of leading digits in naturally occurring numerical datasets, is applied as an exploratory integrity check on the 2022–2024 NewIB-NET dataset in Chapter 9. The rationale for including this analysis is that systematic deviations from Benford's distribution may indicate unusual reporting patterns, which can serve as an additional diagnostic lens alongside the core anomaly detection framework.

The procedure implemented in Python involved extracting the first significant digit of each raw input value across all utilities, computing its empirical frequency distribution, and comparing it to the theoretical Benford distribution using a chi-square goodness-of-fit test. A minimum threshold of ten values per utility is set to ensure robustness, and utilities with insufficient data are excluded from the statistical test. The script further aggregated digit distributions across utilities, plotted the average against the Benford curve, and highlighted the spread between the 10th and 90th percentiles to capture variability.

The results suggest that, while the dataset broadly follows Benford's expectations, deviations are present for digits 6–9, and a subset of utilities did not conform at the 5% significance level.

These outcomes do not necessarily imply manipulation; deviations may reflect sector-specific practices, data entry formats, or genuine operational differences. However, the exercise demonstrates that Benford's Law can complement existing plausibility checks by highlighting cases that warrant closer expert review. In this way, such statistical irregularity tests may provide a low-cost, easily automated addition to broader validation and benchmarking efforts.

Beyond anomaly flagging, observed deviations may also signal underlying issues in measurement or operational management at the utility level. For example, inconsistencies in metering, billing, or water volume tracking. Identifying such patterns could therefore assist utilities not only in improving data quality but also in diagnosing potential inefficiencies or systemic faults in their operations.

B.4 Framework Extensions

Following the completion of this work, numerous possible extensions are identified, reflecting the broad opportunities for further development across all dimensions of the project. These can be grouped into five categories: (1) Data Preparation & Validation, (2) Context-Aware Anomaly Modelling Frameworks, (3) Severity Scoring & Decision Framework, (4) Technical Validation & Evaluation, and (5) Ethical, Political, and Sectoral Considerations. Additional out-of-scope extensions that do not directly align with the core research are discussed in Appendix B.5.

B.4.1 Data Preparation & Validation

1. **Standardised Placeholder Detection:** Current handling of placeholders relies on dataset-specific proxies[123]. While effective for NewIBNET, this approach risks misclassifying legitimate entries and is unlikely to generalise across contexts. Aligning detection with other established survey conventions could create a more robust and transferable framework for identifying placeholder artifacts.
2. **Indicator-Specific k Selection in k-NN Imputation:** The sensitivity analysis shows that the optimal number of neighbours (k) may differ by indicator. Future iterations could adapt k on an indicator-by-indicator basis.
3. **Integration of Industry Thresholds and Expert Knowledge:** Logical validation rules currently enforce general mathematical bounds[124]. Extending these with industry-specific thresholds or expert-defined plausibility ranges could strengthen detection of unrealistic values before reaching the next stage.
4. **Comparative Evaluation of Alternative Imputation Methods:** Chapter 5 compared median and k-NN imputation. Extending this to other approaches could reveal trade-offs between computational complexity, interpretability, and accu-

[123] For example, '1.0' for unknown values

[124] For example: non-negativity and 0–100%

racy, especially for indicators with multimodal or highly skewed distributions.

B.4.2 Context-Aware Anomaly Modelling Frameworks

1. **Clustering for Comparator Discovery:** Instead of relying solely on predefined static categories, unsupervised clustering techniques could be applied to discover "natural" peer groups based on multidimensional characteristics. This may surface latent comparator structures that better capture operational realities than externally imposed thresholds.
2. **Dynamic Comparator Logic Using Macro-Level Indicators:** Comparator assignments could be made adaptive by incorporating macro-level variables such as GDP per capita, urbanisation rates, or climate stress factors. By embedding external metadata, comparator logic could become more context-sensitive.
3. **Temporal Comparators Across Reporting Cycles:** Comparator baselines could be extended longitudinally by incorporating prior submissions of the same utility or region. This would create comparators that evolve over time, flagging not only cross-sectional anomalies but also implausible temporal shifts.

B.4.3 Severity Scoring & Decision Framework

1. **Alternative Deviation Metrics Beyond Z-Scores:** While z-scores offer a simple, interpretable basis for severity scoring, they assume approximate normality and may perform poorly in sparse, skewed, or multimodal datasets. Extending the framework to include alternative metrics – such as Chi-square tests for categorical indicators or robust non-parametric measures – could improve reliability in heterogeneous contexts.
2. **Exploring Alternative Weighting Schemes:** The current framework applies a variance-based weighting approach to balance comparator influence. Future work could trial alternative schemes – such as entropy-based, equal, or expert-informed weights – to evaluate whether different methods yield more stable, interpretable, or context-appropriate results across indicators.
3. **Adaptive Threshold Calibration:** The current severity tiers are anchored in fixed statistical cut-offs. Future work could explore adaptive thresholding, where cut-offs are adjusted dynamically based on indicator-specific distributions, sample

sizes, or historical performance variability. This could mitigate risks of over-flagging in indicators with naturally high dispersion or under-flagging in those with tightly constrained ranges.

4. **Utility-Level Composite Scoring:** The present framework operates at the indicator level. Developing an aggregated severity index per utility – combining across indicators with weights reflecting thematic or operational priorities – could support higher-level decision-making, helping reviewers prioritise which utilities require the most urgent follow-up.
5. **Risk-Based Severity Prioritisation:** Not all anomalies are equally consequential. Future extensions could incorporate impact-based weighting, where severity is scaled not just by statistical extremity but also by the potential operational, financial, or public health risks associated with each indicator. For example, deviations in water quality may warrant greater urgency than anomalies in employee ratios, even if statistically similar.
6. **Communication and Interpretability of Severity Flags:** Beyond algorithmic improvements, extensions could focus on how severity scores are communicated to end-users. Visual dashboards, plain-language summaries, or confidence intervals could help reviewers and utilities better understand the rationale behind each flag, increasing trust and uptake of the system.

B.4.4 Technical Validation & Evaluation

1. **Cross-Dataset Integration and Comparison:** Merging the Indonesian PERPAMSI dataset with the NewIBNET dataset could allow testing how the system performs on a combined sample. This would assess the robustness of comparators and weighting logic when applied across heterogeneous but partially overlapping contexts.
2. **Comparator Removal Experiments:** Running NewIBNET analyses without regional or income comparators – mirroring the PERPAMSI structure – could clarify how much these dimensions contribute to or distort severity scoring. Comparing outcomes across reduced and full models would shed light on the marginal value of different comparator layers.
3. **Broader Utility Datasets Beyond NewIBNET:** Future work could extend validation to additional utility datasets beyond

NewIBNET and PERPAMSI. This would test transferability and highlight whether anomaly detection logic generalises across institutional and geographic settings.

4. **Stress-Testing Under Synthetic Data Scenarios:** Simulated datasets with controlled anomalies could be used to benchmark how consistently the system detects known issues.

B.4.5 Ethical, Political, and Sectoral Considerations

1. **Counterfactual Bias Testing:** Conduct "what if" tests by re-assigning utilities into different comparator groups to see if anomalies persist. This helps determine whether anomaly flags are a product of real performance differences or artefacts of classification choices.
2. **Framing & Language Sensitivity:** Testing alternative framings with the team and utility participants could reveal how linguistic framing affects utility engagement and willingness to improve.
3. **Regulatory Data Cross-Checking:** Cross-checking regulatory online data can provide external validation or supplementary evidence for reported utility data, reducing dependency on self-reported submissions.

B.5 Out-of-Scope Extensions

The proposed framework developed in this thesis represents an initial step toward automation, but several targeted enhancements fell outside its immediate scope due to time constraints, institutional boundaries, or infrastructural limitations. These are technical features or potential upstream adjustments that directly emerged from system analysis.

First, the current survey form lacks hardcoded input constraints, allowing implausible values[125]. Introducing guardrails co-designed with sector experts would help preserve input realism and reduce downstream anomalies.

Second, flagging history is not systematically tracked, limiting the ability to assess alignment between automated and manual reviews. Building this into future data cycles would provide the labelled data necessary for more advanced methods, including supervised machine learning.

Third, inconsistencies in utility naming and year-to-year submissions complicate longitudinal analysis. A unified utility reg-

[125] For example, a population service size of 2 billion, seen in Chapter 5

istry with standardised identifiers could strengthen data coherence and comparability.

Finally, technical integration remains limited: survey data is still handled manually through Excel, rather than via a centralised database. Direct integration would streamline workflows and allow the flagging system to adapt dynamically to real-time feedback.

Together, these extensions represent natural growth areas for NewIBNET, pointing toward a more scalable and cohesive validation ecosystem.

C
Validation & Evaluation

This appendix chapter provides extended material on the validation and evaluation of the proposed system, complementing the discussions in Chapters 8 and 9. It includes a detailed presentation of the PERPAMSI dataset results, the full set of expert survey questions, and additional analysis of flagging patterns across utility groups.

C.1 PERPAMSI Dataset Results

This section presents the application of the proposed flagging framework to the PERPAMSI dataset, comprising 398 utility records. For privacy reasons, utility names and identifying details are not disclosed.

C.1.1 Mapping Process

The original dataset is structured in Bahasa Indonesia. To align it with NewIBNET's raw input titles, the columns are translated into English using Google Translate[126] and mapped accordingly. The mapping below specifies the column number, English translation, and original Bahasa Indonesia wording.

An important factor to consider here is that not all raw inputs and indicators could be mapped from PERPAMSI to NewIBNET. A fairer comparison could involve consultation with a PERPAMSI expert to determine whether a more accurate mapping could be established, enabling a fully functioning framework.

- #3 Company Name "Nama Perusahaan"

[126] **Google Translate** is a free online tool by Google that automatically translates text, speech, and websites between multiple languages. See more: https://translate.google.com/

- #137 Population in Administrative Area (persons) “Jumlah Penduduk di Wilayah Administrasi (Jiwa)”
- #138 Population in Service Area (persons) “Jumlah Penduduk di Wilayah Pelayanan (Jiwa)”
- #38 Average Daily Supply (hours/year) “Waktu distributor air ke pelanggan dalam 1 tahun (Jam)”
- #33 Actual Production Volume (m^3) “Volume produksi Riil (m3)”
- #35 Distribution Loss (NRW) “Kehilangan Distribusi (NRW Distribusi)”
- #25 Active Customers This Year “Jumlah Pelanggan Aktif Tahun ini (SL)”
- #133 Total Revenue (Rp) “Total Pendapatan (Rp)”
- #19 Total Water Bill Receipts (Rp) “Jumlah Penerimaan Rekening Air (Rp)”
- #28 Total Complaints “Jumlah Pengaduan”
- #27 Complaints Resolved “Jumlah Pengaduan Selesai Ditangani”
- #30 Total Required Tests “Jumlah yang di wajib di uji”
- #29 Quality Tests Meeting Standards “Jumlah uji kualitas yang memenuhi syarat”
- #15 Operating Expenses (Rp) “Beban Operasi (Rp)”
- #44 Total Employees (persons) “Jumlah Pegawai (orang)”

C.1.2 Flagging Results

The results in Figure C.1 are presented following the same three-stage structure as the framework.

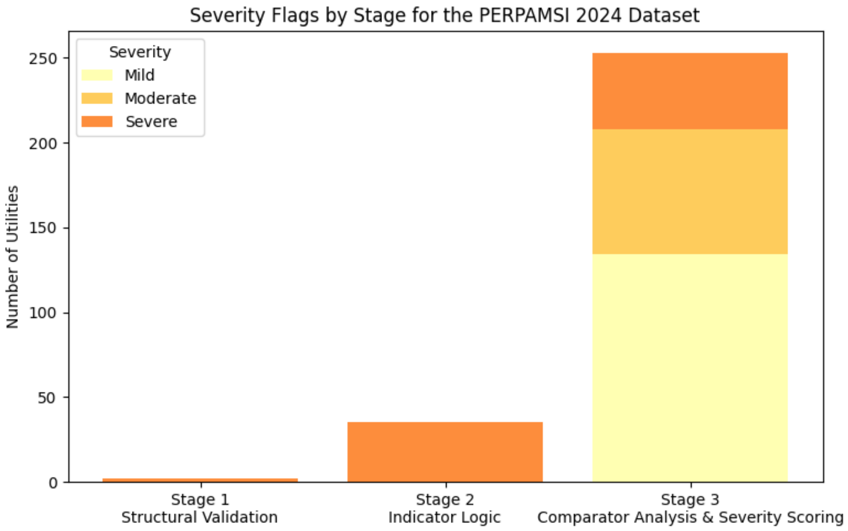


Figure C.1: *Severity Flags by Stage for the PERPAMSI 2024 Dataset*

Stage 1: Structural Validation

2 utilities are flagged for mandatory missing values. The missing values occur in the Annual Water Sales Volume field and the Quality Tests Meeting Standards field.

Stage 2: Indicator Logic

All 35 utilities flagged exhibited the same error: service population size exceeded the total population.

Stage 3: Comparator Analysis and Severity Scoring

There are 136 utilities with *Mild* instances, 73 with *Moderate* instances, and 47 with *Severe* instances (13 utilities overlapping). Further investigation into the causes of the severe cases may be warranted. It is important to note that some instances are repeated; therefore, the number of unique utilities flagged is 207 out of the 398 reported.

Range Comparison

The distribution of weighted z-scores in PERPAMSI and NewIBNET shows broadly similar patterns, with both datasets having negative medians (≈–2.1) and averages close to zero, indicating that most deviations cluster around expected values. PERPAMSI exhibits a wider negative range, while NewIBNET has a slightly higher positive extreme. This suggests that while both systems capture comparable anomaly dynamics, PERPAMSI contains deeper negative outliers, whereas NewIBNET shows a marginally greater spread on the positive side.

A full overview of the descriptive metrics for the weighted z-scores is presented in Table C.1.

Table C.1: *Comparison of Descriptive Metrics for Weighted Z-Scores between NewIBNET and PERPAMSI*

Metric	NewIBNET	PERPAMSI
Minimum	-8.484	-14.383
Maximum	24.458	21.775
Median	-2.134	-2.12
Average	-0.485	-0.372

C.1.3 Indicator Weights

Similar to the NewIBNET dataset, a separate analysis is conducted to examine indicator weights for the PERPAMSI dataset, though the results in Table C.2 show notable differences.

Table C.2: *Covariance-Sum Normalised Weights per Indicator-Comparator for PERPAMSI*

Indicator	Global	Population	Connections
I1	0.2664	0.4645	0.2691
I2	0.2586	0.4576	0.2839
I4	0.2665	0.4616	0.2719
I5	0.2480	0.5026	0.2494
I9	0.2636	0.4574	0.2790
I11	0.2289	0.4874	0.2837
I12	0.2545	0.4118	0.3337
I13	0.2406	0.3902	0.3692

Across all indicators, the Population comparator consistently carries the highest weight (≈0.39–0.50), making it the most influential reference group in the covariance-based framework, while the Global and Connections comparators play relatively smaller and more balanced roles.

In contrast to NewIBNET, the PERPAMSI dataset exhibits a single dominant comparator – Population – since Region and Income are no longer relevant when all utilities belong to the same region and income group. A more balanced benchmark might therefore emerge by re-estimating[127] the NewIBNET weights without Region and Income comparators, enabling a fairer comparison between the two datasets.

C.1.4 Group Distributions

A group analysis is conducted for PERPAMSI shown in Table C.3.

[127] See more possible extensions in Appendix B.4

Table C.3: *Utility Distribution by Category for PERPAMSI*

Category	Population	Connections
Very Low	55	123
Low	213	203
Medium	75	37
High	51	25
Very High	1	6

Similar to NewIBNET, the group sizes are unbalanced, with most utilities concentrated in the *Very Low* and *Low* categories and only a handful in the *Very High* group. This skew highlights the importance of weighting and variance adjustment in the final deviation analysis.

C.2 Expert Survey 2025

The survey, distributed at the end of July 2025 to NewIBNET experts, comprised several sections detailed below. In line with Delft University of Technology Human Research Ethics guidelines[128], the survey is approved by the university and administered via Microsoft Forms[129].

C.2.1 Introduction Page

The survey begins with the statement: *"It will take approximately 10 minutes to complete. All responses will be kept fully anonymous. No identities will be revealed. There is an option to share anonymous quotes."*. And then proceeds to ask the question:

> May answers submitted in this survey be used anonymous results/quotes in the final report?

C.2.2 Case Reflection: Part 1

This section first allows the expert to manually review two utility cases without automation guidance, providing an overview of the utilities as presented below.

[128] **Human Research Ethics:** Refer to: https://www.tudelft.nl/over-tu-delft/strategie/integriteitsbeleid/human-research-ethics

[129] **Microsoft Forms:** An online tool for creating and distributing surveys and collecting responses. See more: https://forms.office.com/

> Case 1: **Would you flag Utility A?** Please provide your reasoning.

> **Utility A:**
> *Country:* Singapore
> *Total Population Service Size:* N/A
> *Average Daily Supply:* 1 hour
> *Total Population in Wastewater Service Area:* 14000
> *Total Service Population in Wastewater Service Area:* 7000
> *Total Wastewater Service Connections:* N/A
> *Total Direct Wastewater Connections:* N/A

> Case 2: **Would you flag Utility B?** Please provide your reasoning.

> **Utility B:**
> *Country:* The Netherlands
> *Total Population Service Size:* 80000
> *Total Operating Expenses:* 1 million (in euros)
> *Total Revenue Collected:* 300k (in euros)

C.2.3 Case Reflection: Part 2

After the manual review responses, the automation results are presented to the expert, accompanied by the statement: "*Based on the previous cases - shown again here - the current automated flagging system flagged the first utility but did not flag the second utility. The reasons for flagging are indicated with* (*FLAGGED - reasoning*) *below:* ". Experts are subsequently presented with:

> Do you agree with the decision **to flag Utility A**? If not, please explain why. If you do agree, but for a different reason than the system's logic, feel free to clarify as well.

> **Utility A:**
> *Country:* Singapore
> *Total Population Service Size:* N/A
> **(FLAGGED - ”missing value detected”)**
> *Average Daily Supply:* 1 hour
> **(FLAGGED - ”placeholder value of 1.0”)**
> *Total Population in Wastewater Service Area:* 14000
> *Total Service Population in Wastewater Service Area:* 7000
> *Total Wastewater Service Connections:* N/A
> **(IMPUTATION - ”value not flagged but filled in using an algorithm”)**
> *Total Direct Wastewater Connections:* N/A
> **(IMPUTATION - ”value not flagged but filled in using an algorithm”)**

> How would you **communicate this back to the utility**?

> Wastewater and sanitation data is optional during survey input, but its absence creates major gaps in later analysis and reduces the reliability of checks. To address this, an algorithm was used to estimate missing wastewater values for select utilities (purely for internal flagging purposes). This has been done for Utility A as seen by: (IMPUTATION - ”value not flagged but filled in using an algorithm”). **Do you agree with this approach? If so, please explain why. If not, feel free to share any questions or concerns.**

> Do you agree with the decision **to not flag Utility B**? If not, please explain why. If you do agree, but for a different reason than the system’s logic, feel free to clarify as well.

> **Utility B:**
> *Country:* The Netherlands
> *Total Population Service Size:* 80000
> *Total Operating Expenses:* 1 million (in euros)
> *Total Revenue Collected:* 300k (in euros)
> **(No flags detected)**

C.2.4 Flag Interpretation & Trust

The next section explores broader flag interpretation, beginning with the prompt:

> **Overview of the Automated System Prototype:**
> The flagging system, inspired by the current reviewer process, reduces manual review time from hours to under 2 minutes. It follows three steps:
>
> - Identifying missing or placeholder data (e.g., missing service population),
> - Validating indicator logic and ranges (e.g., % female employees ≤ 100%),
> - Comparing values against other utilities using weighted deviations (e.g., global averages weighted less if highly variable) within 5 comparator groups (Global Average, Population Service Size, Number of Connections, World Bank Region, and World Bank Income).
>
> The system outputs three types of flags - one per step - with the final stage assigning severity levels: mild, moderate, or severe deviation.

It opens with the questions:

> After reviewing the cases and reading a brief summary of the method, **would you use this automated flagging system?** Please explain your answer.

> Do you think having different severity levels (e.g., mild/moderate/severe based on deviations in step 3) is helpful? Yes/No.

> Could you briefly explain why you selected Yes or No in the question above?

C.2.5 Future Direction & Use

This section concludes by focusing on future steps and feedback on the current system.

> Should **historical consistency** (e.g., comparing this year's submission to previous years for the same utility) be part of future checks?

> Do you see this system being used **beyond anomaly flagging**? If yes, for which purposes?
>
> - Country/Region-specific reporting
> - Benchmarking dashboards
> - Trend analysis
> - Performance alerts
> - Capacity building support
> - None of the above
> - Other

> Are there any other features you would like to see added? Or concerns you'd like to raise?

> Extra Comments.

C.3 Flagging Patterns

A more detailed analysis of the patterns and potential groupings observed during **Stage 3 Comparator Analysis and Severity Scoring** is provided here.

The first comparator examined in greater detail is Population, as shown in Table C.4.

Table C.4: *Utilities per Population Category and Stage 3 Flagging*

Population Category	Total Utilities	Flagged in Stage 3	Proportion (%)
Very Low	104	71	68.27
Low	115	88	76.52
Medium	25	23	92.00
High	29	22	75.86
Very High	16	13	81.25

It can be seen that proportionally, the *Medium* category receives the highest share of flags, with 92% of utilities having at least one type of flag (*Mild*, *Moderate*, *Severe*). This is followed by *Very High*, then *Low* and *High*. The relatively small size of the *Medium* group with 25 utilities may amplify the effect of individual anomalies, driving up the proportion of flagged cases. Similarly, the high proportion in the *Very High* group may partly reflect the limited number of utilities as well, where even a few deviations significantly shift percentages. By contrast, the larger *Low* and *Very Low* categories display lower proportions, though in absolute terms they still account for the majority of flagged utilities. These patterns illustrate the importance of considering group size and variance when interpreting flagging outcomes, as smaller categories may appear disproportionately problematic despite limited underlying evidence.

The second comparator examined in greater detail is Connections, as shown in Table C.5.

Table C.5: *Utilities per Connections Category and Stage 3 Flagging*

Connections Category	Total Utilities	Flagged in Stage 3	Proportion (%)
Very Low	65	45	69.23
Low	132	102	77.27
Medium	39	36	92.31
High	11	10	90.91
Very High	28	23	82.14

A similar trend appears in the Connections categories, with *Medium* and *High* again showing the highest proportional flagging.

The third comparator examined in greater detail is Region, as shown in Table C.6.

Table C.6: *Utilities per Region and Stage 3 Flagging*

Region	Total Utilities	Flagged in Stage 3	Proportion (%)
East Asia and Pacific	25	22	88.00
Europe and Central Asia	40	31	77.50
Latin America and the Caribbean	14	7	50.00
Middle East and North Africa	7	5	71.43
North America	1	0	0.00
South Asia	37	22	59.46
Sub-Saharan Africa	161	129	80.12

For Region, the highest proportional flagging occurs in *East Asia and Pacific*, followed by *Sub-Saharan Africa*. *Europe and Central Asia* also shows a relatively high rate, while *South Asia* and *Latin America* display more moderate levels. These results highlight how regional proportions can be strongly influenced by group size: smaller categories like *East Asia and Pacific* or *Middle East and North Africa* may show high percentages from only a few anomalies, whereas larger groups such as *Sub-Saharan Africa* provide a more stable reflection of underlying data patterns.

The final comparator examined in greater detail is Region, as shown in Table C.7.

Table C.7: *Utilities per Income Level and Stage 3 Flagging*

Income Level	**Total Utilities**	**Flagged in Stage 3**	**Proportion (%)**
High Income	21	14	66.67
Upper-Middle Income	40	31	77.50
Lower-Middle Income	137	99	72.26
Low Income	88	72	81.82

Across income levels, the highest proportional flagging is found among *Low* income utilities, followed by *Upper-Middle* income. *Lower-Middle* and *High* income groups show somewhat lower rates, though still above two-thirds. These results suggest that while data quality challenges occur across all income levels, lower-income utilities may face particular difficulties in consistent reporting, potentially reflecting both capacity constraints and systemic challenges.

D
Process Details

This chapter provides a detailed overview of the thesis development timeline, including the process layout and a meeting and calendar summary of the seven-month trajectory. It also includes a section on practical considerations submitted to the World Bank to support the continuation and integration of their flagging system into practice.

D.1 Development Timeline

The thesis process is structured into three main phases, as shown in Figure D.1.

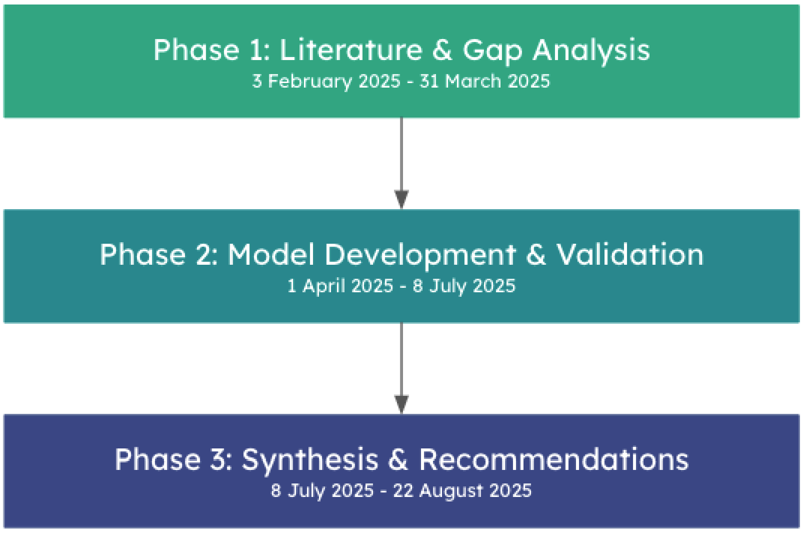

Phase 1: Literature & Gap Analysis
3 February 2025 - 31 March 2025
Phase 2: Model Development & Validation
1 April 2025 - 8 July 2025
Phase 3: Synthesis & Recommendations
8 July 2025 - 22 August 2025

Figure D.1: *Timeline of the thesis development, outlining Phase 1: Literature Review & Gap Analysis, Phase 2: Model Development & Validation, and Phase 3: Synthesis & Recommendations.*

A visual overview of the meetings, deadlines, and phases is shown in Figure D.2. In addition, a separate file is maintained documenting meeting details, including attendance, discussion points, and agreed next steps.

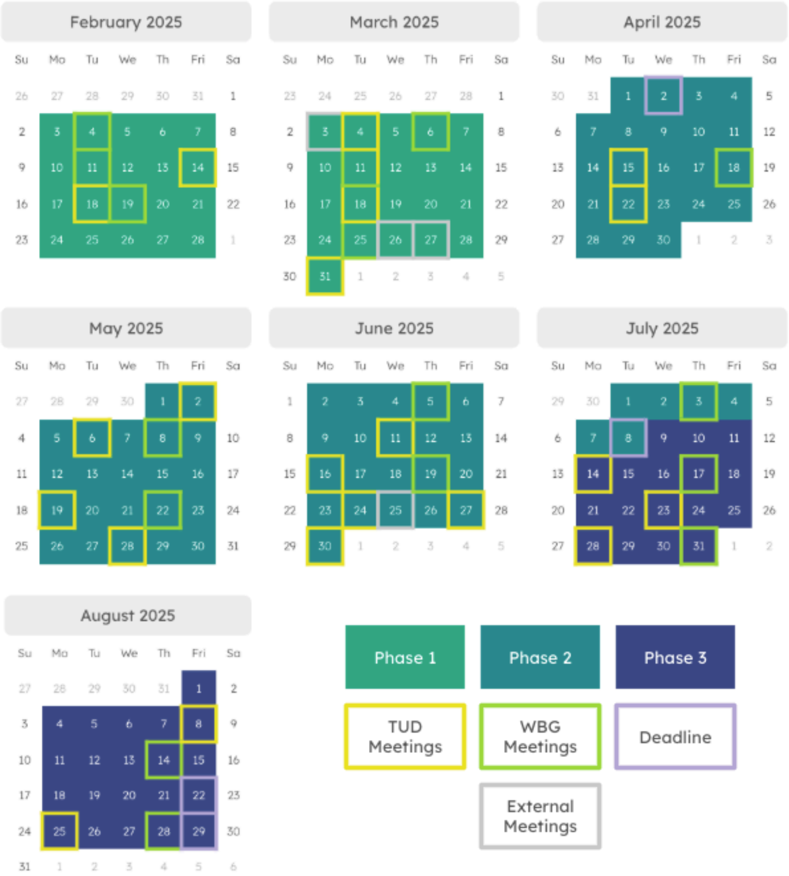


Figure D.2: *This figure presents a calendar overview of the seven-month trajectory, highlighting key meetings, deadlines, and phases.*

D.2 World Bank Manual

This thesis originated from the real-life assignment of developing an optimised flagging system within the NewIBNET framework. In addition to this report, a condensed manual is prepared for reviewers who wish to apply the framework in practice.

No user interface was developed, as this fell outside the thesis scope and timeline. Instead, the NewIBNET team will receive the Python files, accompanied by a manual explaining the required directory structure and basic steps. Reviewers only need to upload the dataset and run the designated files, after which the system automatically generates outputs for flagging and review.

As noted in Appendix B.5, a logical next step would be to work with the World Bank's IT team to integrate the tool directly into the online NewIBNET platform, enabling reviewers to access the system seamlessly and in real time.

All code files and raw input utility data remain private and will not be made publicly available, in line with the confidentiality agreement established with the NewIBNET team.

D.3 Large Language Model Acknowledgement

Reference: OpenAI. (2025). ChatGPT (Feb-Aug 2025 version) [Large Language Model]. https://chat.openai.com/chat. Prompt: "Rephrase: ... [insert sentence]".