Improvements to Standards-Driven DGA Condition Monitoring Methodologies of Oil-Immersed Transformers

Michael Hosseini

A Thesis Submitted for the Degree of Doctor of Philosophy to the Department of Electronic and Electrical Engineering

University of Strathclyde

October 2024

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Abstract

This thesis focusses on advancing a Standard-driven methodology for interpreting *Dissolved Gas Analysis* (DGA) for condition monitoring of oil-immersed *Transformers* (TXs). The focus is on analysing and evaluating the methodology outlined in IEEE C57.104-2019, which differs substantially from both its predecessor and its closest peer: IEC 60599:2022. The implications of these changes are difficult to intuit. This thesis details a comparison of their relative behaviours via case studies using real TX DGA data.

More generally, it can be unclear how to proceed when facing issues attempting a practical deployment. Modifications to a prescribed methodology can undermine the basis of its original validity, however, capturing the original intent can be a time-consuming and nebulous task. This thesis highlights potential barriers and presents relevant quantifiable and analytical advances to facilitate the deployment of the IEEE C57.104-2019 methodology in an automated setting. The rationale is based on a holistic overview of the topic area and on findings from real TX DGA case studies.

The presented improvements to the methodology target the default limits as well as the derivations to both the input and output metrics. These improve the methodology's noise tolerance, metric consistency, and output granularity, respectively. The proposals are intended to preserve the methodology's perceived original intent whilst improving upon its provided decision-support for the screening of TXs using DGA in practical deployment.

Additionally, the thesis explores extending the methodology to incorporate a measure of uncertainty. Though IEEE C57.104-2019 emphasises the importance of uncertainty, it provides limited guidance on its practical application. A novel methodology, grounded in a Standards-based literature review, is proposed to address this gap by quantitatively assessing measurement uncertainty via the propagation of probability distributions. The methodology uses a Monte-Carlo technique, which is validated and demonstrated as a scalable, simple-to-apply solution for larger datasets.

Acknowledgements

In memory of my father, and with thanks to my wife.

Though you did not meet, you together complete my life.

Copyright Acknowledgements

Table of Contents

Table of Abbreviations

Abbreviation	Full Term – General Scope
ASTM	American Society for Testing and Materials
CB	Circuit-Breakers
CBA	Cost-Benefit Analysis
CBM	Condition-Based Maintenance
CD	Critical Difference
CDF	Cumulative Distribution Function
CIPM	International Committee for Weights and Measures
CMA	Condition Monitoring and Assessment
CoF	Consequence of Failure
DETC	De-Energised Tap-Changer
DGA	Dissolved Gas Analysis
EM	Electromagnetic
EoL	End of Life
FBM	Frequency-Based Maintenance
FE	Fixed-Effect
FM	Failure Mode
FMEA	Failure Modes and Effects Analysis
FTA	Fault Tree Analysis
GBN	Gaussian Bayesian Network
GIOS	Gas-in-Oil Standard
GSU TX	Generator Step-Up Transformer
GUF	GUM (*see above*) Uncertainty Framework
GUM	Guide to the Expression of Uncertainty in Measurement
IED	Intelligent Electronic Device
JCGM	Joint Committee for Guides in Metrology
LoD	Limit of Detection
LoQ	Limit of Quantification
LPU	Law of Propagation of Uncertainty
LSA	Lapworth Scoring Algorithm
MAPE	Mean Absolute Percentage Error
MCM	Monte Carlo Method
MSA	Measurement System Analysis
NEI	Normalised Energy Intensity
OEM	Original Equipment Manufacturer
OLCM	On-Line Condition Monitoring
OLDGA	Online Dissolved Gas Analysis
OLS	Ordinary Least Squares
OLTC	On-Line Tap-Changer

OOD	Out-of-Distribution
PFM	Performance Focused Maintenance
PoF	Probability of Failure
RBM	Risk-Based Maintenance
RCM	Reliability Centred Maintenance
RM	Reference Material
RRT	Round Robin Test
RSE	Residual Standard Error
RUL	Remaining Useful Life
S	Analytical Detection Limit (IEC 60599), Limit of Detection
SoH	State of Health
SOP	Standard Operating Procedure
SSE	Sum or Squared Errors
SST	Sum or Squared Total
TAI	Transformer Assessment Index
TBCM	Time-Based Condition Monitoring
TBM	Time-Based Maintenance
TCG	Total Combustible Gas
TDCG	Total Dissolved Combustible Gas
TICM	Transformer Intelligent Condition Monitoring
TX	Transformers
UHF	Ultra-High Frequency
VIM	International Vocabulary of Metrology
R	Reproducibility (IEC 60567) / Reproducibility Limit (ISO 5725)
r	Repeatability (IEC 60567) / Repeatability Limit (ISO 5725)

Abbreviation	**Full Term – Fault Types**
C	Hot Spots with Paper Carbonisation (Temperatures < 200 °C)
D	Arcing-Related Faults (General)
D1	Low Energy Electrical Discharges
D2	High Energy Electrical Discharges
DT	Mixed Electrical and Thermal Faults
O	Overheating (Temperatures < 250 °C)
PD	Partial Discharge
S	Stray Gassing of Oil (Temperatures < 200 °C)
T	Thermal-Related Fault (General)
T1	Thermal Faults (Temperatures < 300 °C)
T2	Thermal Faults (300 °C < Temperatures < 700 °C)
T3	Thermal Faults (700 °C < Temperatures)

Abbreviation	**Full Term – Gases**
C_2H_2	Acetylene
C_2H_4	Ethylene

C_2H_6	Ethane
CH_4	Methane
CO	Carbon Monoxide
CO_2	Carbon Dioxide
H_2	Hydrogen
H_2O	Water
N_2	Nitrogen
O_2	Oxygen
Abbreviation	**Full Term – IEEE C57.104–2019 Terms**
Ł1	DGA Status 1 (~'probably normal')
Ł2	DGA Status 2 (~possibly abnormal')
Ł3	DGA Status 3 (~'probably abnormal')
L1	DGA Status 1 for a single gas
L2	DGA Status 2 for a single gas
L3	DGA Status 3 for a single gas
$\mathbb{T}1$	Limit Table 1 (90^{th} % of absolute gas levels) [PPM]
$\mathbb{T}2$	Limit Table 2 (95^{th} % of absolute gas levels) [PPM]
$\mathbb{T}3$	Limit Table 3 (95^{th} % of difference in gas levels) [ΔPPM]
$\mathbb{T}4$	Table 4 (95^{th} % of average gassing rate) [ΔPPM /Year]
S1	Protocol when only $\mathbb{T}1$ and $\mathbb{T}2$ are applicable
S2	Protocol when only $\mathbb{T}1$, $\mathbb{T}2$, and $\mathbb{T}3$ are applicable
S3	Protocol when all Limit Tables, $\mathbb{T}1$–4, are applicable
Abbreviation	**Full Term – IEC 60599:2022 Terms**
Ł1	Typical Condition
Ł2	Alert Condition
Ł3	Alarm Condition
L1	Typical Condition for a single gas
L2	Alert Condition for a single gas
L3	Alarm Condition for a single gas
L_1	Typical Gas Concentration Limit [PPM]
L_2	Alarm Gas Concentration Limit [PPM]
G_1	Typical Gassing Rate Limit [ΔPPM /Year]
G_2	Alarm Gassing Rate Limit [ΔPPM /Year]

Words *Italicised* in thesis are intended in a more formal, defined context relevant to the pertaining literature than they may otherwise be commonly used.

Table of Symbols

Symbol	Remarks – Standards-Driven Uncertainty Literature Review
$\mathcal{N}_y$	Gaussian Distribution describing y
μ	'True' Value, (ISO 5725 allows this to be a *Reference Material*)
$\hat{\mu}$	Measured, or Estimated 'True' Value
m	Mean of all Measurements, used as estimated μ (ISO 5725)
δ	Bias (Trueness if pertaining to Accuracy)
$\hat{\delta}$	Estimated Bias (Trueness if pertaining to Accuracy) (ISO 21748)
B	Laboratory Component of Bias (ISO 21748)
Δ_y	Laboratory Bias (ISO 21748)
e	Random Error under Repeatability Conditions
s_y	Estimated Standard Deviation of y based on Measurements
σ_y	Standard Deviation of y, (ISO 5725 allows an estimate for this)
$u(y)$	Standard Uncertainty Associated with Value, y (ISO/IEC 98-3)
L	Inter-Laboratory Component of Uncertainty (ISO 5725)
s_L	Estimated σ_L of B
W	Within Laboratory Component of Uncertainty (IEC 60567)
s_W	Estimated σ_W of e
r	Repeatability Component of Uncertainty (ISO 5725)
s_r	Estimated σ_r using averaged variance of e from all laboratories
R	Reproducibility Component of Uncertainty (ISO 5725)
s_R	Estimated σ_R using s_L and s_r
I	Intermediate Precision Component of Uncertainty (ISO 5725)
s_I	Estimated σ_I using s_r and other context-specific components
CD	Critical Distance, interval that measurements should fall within
k	Coverage Factor (ISO/IEC 98-3)
f	*Critical Range Factor* (ISO 5725), similar to *Coverage Factor*
S	Limit of Detection
Symbol	**Remarks – Notation used Sub-Section 4.1.4 Onwards**
X_i	Measured Sample Date, where $i = 1$ is the most recent sample
Y_i	Measured Gas Value, where $i = 1$ is the most recent sample
y_i	Possible value representing Y_i
$\check{y}$, $\bar{y}$, $\hat{y}$	Minimum, Mean, and Maximum of value, Y
N	Oldest applicable Y_i (i.e., maximum value for i)
$\mathbb{N}$	Abstract merging of i, where $i = [3, N]$ (as marginal distribution)
$y_{\mathbb{N}}$	Abstract merging of y_i, where $i = [3, N]$ (as marginal distribution)
β_0	Intercept of Linear Regression
β_1	Slope Coefficient of Linear Regression
k	Coverage Factor, assume $k = 1$ unless indicated otherwise

c_i	Change in Slope Coefficient, β_1, as measurement, y_i, changes
$\delta(\dots)$	Kronecker delta function
$f(\dots)$	Probability Density Function
$F(\dots)$	Cumulative Distribution Function
$*$	Generic Convolution operation
$\circledast$	Chained Convolutions (analogous as Σ is to $+$)
T_i	Limit Tables, $\mathbb{T}\mathbb{1}$–$\mathbb{4}$, **not** being exceeded. $\neg T_i$ is being exceeded.
τ_i	Applicable Limit from Limit Tables, $\mathbb{T}\mathbb{1}$–$\mathbb{4}$
$P(\star)$	Entire attainable Probability Space based on $Y_{i=[1,N]}$
α	Relative Uncertainty factor (to be multiplied by Y_i)
W	Difference between $\check{y}$ and $\hat{y}$ described by a Δ distribution.
$SEM_{\beta 1}$	Standard Error of Mean of β_1
W_P	Confidence Interval for the given probability of success, P
$\mathcal{N}_N$	Described Method: Using $\mathcal{N}$ Distribution, and $Y_{i=[1,N]}$
$\mathcal{N}_{\mathbb{N}}$	Described Method: Using $\mathcal{N}$ Distribution, and $Y_{i=\{1,2,\mathbb{N}\}}$
Δ_N	Described Method: Using Δ Distribution, and $Y_{i=[1,N]}$
$\mathcal{N}^{\Delta}$	Described Method: Using $\mathcal{N}$ approximating a Δ Distribution
Σ	Covariance Matrix for Inter-Gas Relationships
ρ	Assumed-shared Inter-Gas Correlations

1. Introduction and Summary

Thesis Overview

Large *Transformers* (TXs) are expensive assets critical to the electrical infrastructure on a national scale. *Dissolved Gas Analysis* (DGA) of mineral oil within TXs is a well-established condition monitoring technique that has demonstrated its unique value in its ability to reliably detect a large range of developing faults over the decades. However, as domain knowledge and monitoring technologies advance, robust DGA interpretation methodologies for assessing TX condition and ensuring their continued intended function remain as an active area of research. One barrier is the difficulty in amassing sufficient relevant data of failures to develop novel methodologies. There is therefore a demand for Standards-based methodologies developed through the pooling of resources from numerous bodies. Even when novel methodologies are pursued, Standards-based methodologies can contribute by serving as recognised benchmarks. Thus, in representing the current status quo, advancements to Standards-based methodologies constitute significant developments in the field.

This thesis focusses on the interpretation of DGA of mineral oil within TXs for *Condition Monitoring and Assessment* (CMA) screening purposes using Standards-driven methodologies. The reviewed methodologies are interpreted and evaluated for the context of use in an automated implementation. The focus is on the screening methodology outlined in IEEE C57.104-2019, which interprets DGA samples to output a *DGA Status* of 1–3 to indicate whether a TX requires further attention. This improves resource allocation for CMA purposes. The thesis identifies potential weaknesses in the IEEE C57.104-2019 methodology and then contributes proposed improvements to its performance in practical deployment. The contributions address excessive flagging and enhance decision-support by improving its output granularity to further facilitate TX comparisons. Furthermore, the IEEE C57.104-2019 identifies the significance of *Uncertainty* but lacks detailed guidance on its incorporation. Therefore, this thesis contributes a novel methodology for quantitatively integrating *Measurement Uncertainty* into the methodology outlined in IEEE C57.104-2019. This novel methodology can inform engineers when outputs could be explained due to measurement noise and thus inform decision-making.

The scope of this thesis is divided into three *Research Themes*. The TX DGA screening methodology outlined in IEEE C57.104-2019 [1] changed substantially from both its predecessor, IEEE C57.104-2008, and arguably, its closest peer: IEC 60599:2022 [2]. *Research Theme 1A* considers the implications the changes introduced in the new edition of the IEEE C57.104 methodology have on practical deployment. This thesis details a comparison of their relative behaviours via case studies using real TX DGA data to help convey the significance of the changes made with this version of the IEEE C57.104 methodology. *Research Theme 1B* then proposes improvements to said methodology based primarily on the findings from *Research Theme 1A*. Lastly, *Research Theme 2* concentrates on quantifying *Uncertainty* within the IEEE C57.104 methodology to enable increased confidence in DGA-based TX CMA.

Research Theme 1

A challenge to modifying a *Standard Operating Procedure* (SOP) or methodology is that it can undermine the basis of its original validity. However, capturing the original intent can be a time-consuming and nebulous task. It can therefore be unclear how to proceed when facing issues attempting a practical deployment. This can result in the rejection of methodologies for potentially surmountable issues. This thesis offers a holistic overview of the topic for sufficient context, based on the shared bibliography of two of the most widely recognised methodologies currently published on this topic: IEEE C57.104-2019 [1] and IEC 60599:2022 [2]. Therefore, there is a heavy emphasis on the Technical Brochures by industrial groups such as CIGRE, and Standards and Guidance documents by technical groups such as IEC and ISO.

The focus is on the methodology introduced in IEEE C57.104-2019 as it is substantially different to both its predecessor and IEC 60599:2022. Although its output is still primarily based on a comparison of DGA data against limits held across tables as before, the derivations for the input and output metrics differ. The relative behaviours of some other methodologies are compared to provide meaningful context when gauging the impact of these changes. The primary comparison is against IEC 60599:2022 due to its comparable scope and pedigree. The *Normalised Energy Intensity* (NEI) method [1, Annex F], [3], [4] is selected as it is mentioned within IEEE C57.104-2019 as a potential alternative approach. Lastly, the *Lapworth Scoring Algorithm* (LSA) [5] is chosen and

evaluated for comparison against NEI given their similarities and its relevance to industry.

Automated implementations of the *Screening* components of the methodologies are applied to case studies of real TX DGA data to identify and demonstrate potential problems. Based on the findings, several improvements to the methodology outlined in IEEE C57.104-2019 are proposed. The proposals directly improve the reliability and consistency of the TX evaluations provided by the IEEE C57.104-2019 methodology in practical deployment. Preserving the perceived intent of each modified aspect of the original methodology has been prioritised. Similarly, the scope for these improvements is kept to those comparable in complexity and intuitiveness to their original counterparts; the aspiration being that some aspects could be incorporated into future editions of this methodology. A summary of the novel thesis contributions to the field of study is listed:

1. Default limit values in the IEEE C57.104-2019 [1] methodology's Tables 3 and 4: The case studies demonstrated the methodology tended towards excessive flagging, with limit values of zero identified as a key contributor. It is therefore recommended for them to be adjusted where applicable to improve noise tolerance.
2. Derivation of the metric used in the IEEE C57.104-2019 [1] methodology's Table 4: The changes to the metric in Table 4 of the new version of [1] are among the most impactful. This metric is intended to quantify the extent of active gassing. However, the current derivation is demonstrated to have issues when the sampling frequency varies, which would be whenever a problem is suspected. Implementing a minimum timespan as an added stipulation to the derivation would enhance consistency during this critical period.
3. Derivation for both the IEEE methodology's per-gas, and combined, output metrics: These output metrics are intended to convey the extent a TX's DGA results are deemed suspicious. The impact of excessive flagging is shown to be exacerbated by the lack of granularity in the outputs. It is therefore argued that adding granularity will improve decision support by differentiating TXs of otherwise identical output category. To this end, two proposals are made:
   - using a linear interpolation between its Table 1 and Table 2 outputs for the derivation of the per-gas output, and

- using a weighted combination of the per-gas outputs for the combined output metrics.

Additionally, adopting a 0–1 scale for [1]'s outputs would improve compatibility when integrating into a multi-index condition monitoring programme.

Research Theme 2

Uncertainty associated with a measurement can at times explain why a given TX is flagged. Its explicit quantification is necessary to accurately identify these cases to use as a basis for avoiding otherwise costly and unnecessary interventions. Although IEEE C57.104-2019 highlights the importance of *Uncertainty*, it does not provide an explicit means to incorporate it. This is also true for the methodology outlined in IEC 60599:2022. This thesis contributes to this topic by identifying and reviewing relevant Standards-based guidance to gauge their applicability to TX DGA interpretation for screening purposes. Furthermore, barriers to their interpretation and practical implementation are identified and discussed in detail. Then, a Standards-driven methodology is proposed to implement the integration of *Measurement Uncertainty* into the methodology outlined in [1] specifically. The viability of this proposed methodology is demonstrated using two techniques: numerical integration and *Monte-Carlo Method* (MCM). Case studies using real TX DGA data demonstrate its potential to improve decision-support by explicitly conveying the expected impact of *Measurement Uncertainty*.

Thesis Structure

Chapter 2 provides a background on the topics underpinning the rest of the thesis. Section 2.1 covers the specifics of a TX, the expected *Failure Modes* (FMs), and associated symptomatic pre-cursors. Section 2.2 explores the role of a *Screening* output for the *Condition Monitoring and Assessment* (CMA) of TXs. Section 2.3 discusses the concept of *Uncertainty* and its relation to TX CMA, including different potential approaches to incorporate *Uncertainty*. Finally, Section 2.4 outlines the principles of DGA and its interpretation for TX CMA.

Chapter 3 presents a detailed Standards-based literature review focussing on two topics. The first is DGA interpretation for TX CMA, and the second is quantifying *Uncertainty*. Section 3.1 reviews and compares four DGA interpretation methodologies

in detail: IEEE C57.104-2019 [1], IEC 60599:2022 [2], the *Normalised Energy Intensity* (NEI) method [1, Annex F], [3], [4], and the *Lapworth Scoring Algorithm* (LSA) [5]. Section 3.2 then considers the application of *Measurement Uncertainty* for TX DGA CMA. A focus is on the practical implications suggested by the relevant Standards and Guidance documents, such as what information may be expected to be available for the *Screening* analysis, and how they advise its use. Practical challenges to the application of said advice are highlighted and discussed.

Chapter 4 considers the practical implications in more detail. Section 4.1 uses simple conceptual experiments to assess the expected practical impact of these topics on the metrics used within the DGA interpretation methodologies. Section 4.2 then uses automated implementations of the explored DGA interpretation methodologies developed for this thesis for a detailed case-study driven comparative analysis. These implementations are applied to real TX DGA data to explore their relative behaviours and demonstrate the implications of decisions made, such as the metric selection. The findings constitute a conclusion of *Research Theme 1A*.

Chapter 5 uses these findings to improve the methodology outlined in IEEE C57.104-2019 [1]. Section 5.1 concludes *Research Theme 1B* by presenting the numbered contributions above. Sub-Section 5.1.1 and 5.1.2 cover contributions **1** and **2**, improving the default limits and metric definitions used, respectively. The remainder of Section 5.1 covers contribution **3**, improving the output metric derivation on both a per-gas level, and as a combined output. Section 5.2 concludes *Research Theme 2* by presenting the contributions related to quantifying *Measurement Uncertainty* within the IEEE C57.104-2019 methodology. The challenge is mathematically defined in Sub-Section 5.2.1 and demonstrated to be non-trivial to algebraically solve in Sub-Section 5.2.2. Two novel methodologies are presented in Sub-Section 5.2.3 for overcoming this challenge. Sub-Section 5.2.4 presents extensions to one of the novel methodologies, one by extending it into *Diagnostic* stage of the IEEE C57.104-2019 methodology, and the one by integrating inter-gas correlations. Lastly, Sub-Section 5.2.5 uses real TX DGA data to demonstrate the practical deployment of the proposed methodology and its potential impact on TX DGA CMA using the IEEE C57.104-2019 methodology.

Chapter 6 summarises the conclusions and potential avenues for further work.

Thesis Publications

- Industry Reports: University of Strathclyde's ANRC, Project 11-3's reports: 1, 2, 2a, and 3. These were published between 2020–2021. These looked at implementing and comparing the four DGA interpretation methods with an emphasis on *Online DGA* (OLDGA). Additionally, a hybrid combination of the methods was proposed, as well as a novel statistical approach for *Anomaly Detection* using OLDGA.
- Conference Publication: "Construction of a Transformer DGA Health Index Based on DGA Screening Processes", CEIDP, 2020, pages 391-394. This explored some of the proposed modifications to the derivations of the output metrics of [1].
- Conference Publication: "Propagating Uncertainty using IEEE Std C57.104-2019 Dissolved Gas Analysis Methodology for Transformers", ISH, 2023, pages 698-704. This explored the integration of *Measurement Uncertainty* into [1].

Supervisors

Prof. Stephen McArthur and Dr Bruce Stephen. Professor Brian G. Stewart was the Principal Investigator associated with the industry reports and was also instrumental in this thesis. All from the Department of Electronic and Electrical Engineering, University of Strathclyde.

2. Background

Chapter Scope

This Chapter provides contextual background on three topics relevant to the thesis. First, the role a *Screening* output plays within wider *Asset Management*, focussing on those derived from *Dissolved Gas Analysis* (DGA) of an oil-immersed *Transformer* (TX) for *Condition Monitoring and Assessment* (CMA). Second, it covers the concepts of *Uncertainty*, *Conformity*, and unexpected results in relation to TX CMA. Third, the premise of TX DGA is explained. There are numerous terms and concepts in extant literature, which are sometimes used interchangeably, creating potential ambiguity. This Chapter clarifies this thesis's interpretation of these terms as they relate to TX DGA-based CMA.

Chapter Structure

Section 2.1 introduces the components and *Failure Modes* (FMs) associated with mineral-oil immersed TXs. Section 2.2 begins by characterising *Condition Monitoring* approaches, as well aspects influencing the choice of *Condition Monitoring Techniques*. Next, the process of *Condition Assessment*, including the construction of *Transformer Assessment Indices* (TAIs), is conceptually explained. Section 2.2 concludes by discussing how these contribute towards a *Condition Monitoring Programme*, providing brief overviews of example implementations.

Section 2.3 presents an overview of *Uncertainty* for CMA. In this context, metrics are expected to be related to symptomatic poor health in a TX. If a metric is above a limit, action may be taken. However, there is often an implicit or explicit allowance for the possibility that the result is an unrepeatable *Anomaly*. Section 2.3 explores different potential sources and causes of *Uncertainty*, followed by a brief overview of techniques available to incorporate *Uncertainty* into a TAI.

Section 2.4 introduces the premise of DGA and its basic interpretation. This includes a characterisation of typical metrics used for *Screening*. Section 2.4 also briefly discusses the implications of different DGA *Sampling Techniques*, particularly given the increasing prevalence of *On-Line Condition Monitoring DGA* (OLDGA), which will influence the interpretation of DGA.

2.1. Transformer Overview

2.1.1. Network Context

Fig. 2-1 illustrates a simplified model of the electricity network within the UK [6]. The network is segmented by voltage levels, though specific breakpoints differ by region. Throughout these points, there may be electrical supply via electricity generation, and electrical demand via electricity consumption. Higher voltage levels are generally motivated by reduced transmission losses over long distances. The voltage levels are reduced downstream for end-user safety and convenience. It is TXs that connect differing voltage levels: a *Step-Up* or *Step-Down* TX is used to increase or decrease the voltage levels, respectively. A *Generator Step-Up* (GSU) TX is responsible for offloading the electricity from larger electrical generators to the transmission network. This thesis focusses on GSU and Transmission TXs.

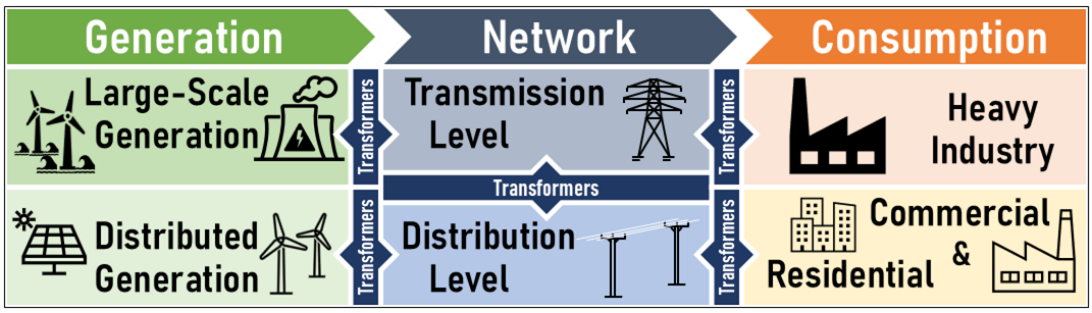
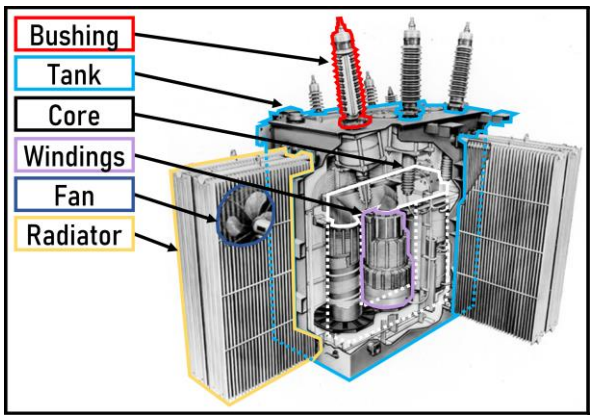


Fig. 2-1: Simplified electricity network in the UK

2.1.2. Components and Subsystems

The design and operating principles of a TX influence the most appropriate *Condition Monitoring Techniques*. However, larger TXs are complex assets with long lifespans, often featuring bespoke designs tailored to specific purchase orders [7, Sec. 7]. The following overview is based on [7, Ch. 8–12], [8, Ch. 4], [9, Ch. 2], [10, Ch. 2–4]; these can be referred to for further details on TX design. A three-phase oil filled TX is shown in Fig. 2-2, although note there is no *Conservator* or visible *Tap-Changer*. The mechanism by which an AC TX transfers power from one electrical circuit to another is via *Electromagnetic* (EM) induction. The electric circuits are linked via a magnetic circuit that is described by the flux flow through the TX's *Core*. The *Turns* ratio between the *Windings* dictate the voltage ratio between circuits.



Source: Modified from [10, Fig. 2.3]

Fig. 2-2: Annotated Cutaway of Transformer

Transformer Components

The specific components depend on the manufacturer and TX type, but the following components are generally present in an oil-filled TX:

Core

The *Core* is traditionally made from high-purity iron, but recent designs may include low-carbon silicon alloys for efficiency. It is formed from laminated sheets to reduce losses via eddy currents [14]. The magnetic circuit primarily travels along the *Core* material and is influenced by its design, which is either a *Shell* or *Core* design. The *Shell* design encases the *Windings* whereas the *Core* design is instead wrapped by them. The former may also require magnetic shielding from the *Tank*.

Windings

The TX *Winding Coils* typically consist of *Turns*, usually wrapped in paper insulation and mounted on a *Winding Mandrel*. Each *Turn* consists of multiple, individually insulated copper strands. The *Coils* are typically either arranged in inter-connected concentric disks, or in a helical pattern. Spacers are installed to facilitate oil flow and ensure electrical separation, typically made using paper or pressboard. There are typically two pairs of *Windings* per phase, representing the two electrical circuits.

Tap-Changers

The *Turns* ratio between the *Windings* dictate the voltage ratio between the circuits, and this can be adjusted by a *Tap-Changer*. An *On-Line Tap-Changer* (OLTC) can do this whilst the TX is on-line whereas a *De-Energised Tap-Changer* (DETC) requires the TX to first be de-energised. A DETC is intended to be used less frequently, such as due

to a change in system configuration. An OLTC is typically housed separately to prevent cross-contamination of oil with the *Tank*, though this is not always the case and is an important detail to ascertain prior to many diagnostic tests.

Bushings

The *Windings* are generally connected to *Bushings*, which provide electrical separation from the *Tank*. These will typically have a weather-shielding skirting that also increases electrical creep distances. Many *Bushings* contain oil insulation that must remain separate from the oil in the *Tank*. *Lightning* (*Surge*) *Arrestors* may be installed to protect against lightning strikes or through-faults.

Tank

The TX is housed within a *Tank*, generally manufactured of hot-rolled, unalloyed steel sheets. This provides shielding from the environment and contains the oil within. The *Tank* may experience significant mechanical stresses from pressure changes and magnetic-induced vibrations of the *Core*.

Internal Insulation

It is essential that the electrical circuits remain electrically insulated. There are therefore multiple applications of insulation within a TX. There are very generally two categories: *Liquid Insulation* and *Solid Insulation*. The *Solid Insulation* is generally constructed of cellulosic materials such as kraft paper or pressboards.

Oil Preservation System

The mineral oil's electrical insulating properties are highly dependent on its purity. TXs will therefore have mechanisms to reduce unwanted ingress of external elements. Changes in the oil volume, primarily driven by changes in its temperature, can cause fluctuations in the internal pressure of tank that could either draw air and moisture in, or potentially push oil out. There are different systems intended to avoid this.

- *Free-Breathing Transformers* – The most common type of TX and does little to isolate the oil. Sometimes a *Breather* is used to maintain dryness of the oil.
- *Sealed Tank Transformers* – A sealed system intended to withstand the expected changes in pressure without any induced ingress, often used for smaller TXs.
- *Pressure Regulated Gas Blanket System* – Nitrogen is used to create positive pressure within the *Tank* to discourage ingress. A control system is used to maintain desired

pressure, drawing in more nitrogen from bottles when needed, and purging excess nitrogen to reduce pressure.

- *Oil Conservator* / *Surge Tank* / *Oil Pillow* – A tank partially filled with insulation oil that can be drawn from, or added to, depending on the *Tank*'s oil volume that is usually included on larger TXs. The tank could be sealed or free breathing. If the latter, it could have a *Diaphragm* or a *Bladder* to keep the air separated from the oil, and potentially a *Breather* to reduce moisture levels.

Cooling Equipment

The mineral oil's electrical insulating properties are also dependent on its temperature; both directly, and indirectly, as elevated temperatures accelerate the oil's degradation. Different cooling solutions are used depending on the size and type of TX. It is typical for larger power TXs to be capable of operating in multiple modes: one as self-cooling, and others with various levels of powered cooling. Some of the common components used to help with the cooling are the following:

- *Radiators* – Dissipates heat into the environment.
- *Fans* – External to the *Tank*, increases air flow across the *Radiators*.
- *Oil Pumps* – Improves the circulation of the oil to improve heat dissipation.
- *Heat Exchangers* – Used to further improve the heat dissipation.

Protection and Monitoring Devices

There are various protection devices such as *Pressure Relief Valves* (PRV) and *Buchholz Relays* that serve as contingency mechanisms to mitigate catastrophic failures. This can include *Fire Suppression Systems*. In addition, there are a suite of monitoring devices used, such as temperature and gas detectors, or oil level indicators.

A PRV can activate very rapidly when its pressure limit is exceeded as an emergency measure [9, Ch. 4]. There may also be a *Sudden Pressure Relay* to detect rapid pressure changes. Its limit would be set below that of the PRV, and it would raise an alarm and can de-energise the TX [9, Ch. 4]. The *Buchholz Relay* is intended for TXs with *Conservators*. It captures gases released due to bubbling, which can then be analysed via DGA. If too much gas has accumulated, it too can activate an alarm and can de-energise the TX [9, Ch. 4].

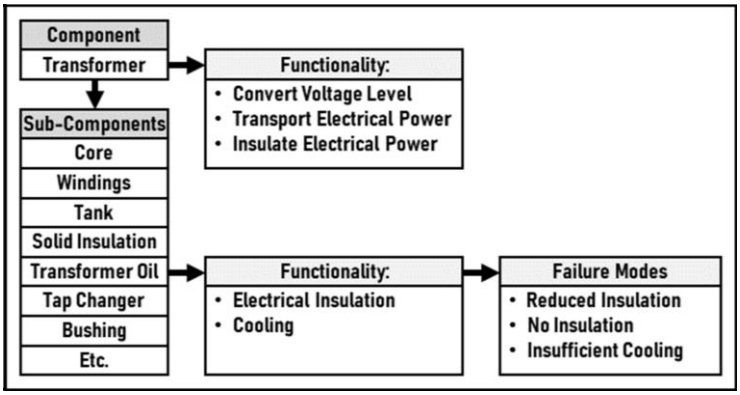
Transformer Subsystems

TXs contain multiple components and potentially coexisting *Failure Modes* (FMs). Grouping components into functional subsystems simplifies *Failure Modes and Effects Analysis* (FMEA) [12, Sec. 5]. Annex C in [13] states that a *Condition Monitoring Strategy* can be devised by analysing each *Functional Component*'s FMs and their corresponding symptoms to select appropriate sensors. As per Annex C of [13], a *Functional Component* is defined as a set of components that have a common functional goal to the TX's operation. This can mean some physical sub-components are present in multiple components if they serve multiple purposes. Different literature will categorise differently. As per [10, Ch. 2], the main function of a TX is to convert electric power from one voltage level to another whilst efficiently conducting high current levels and effectively withstanding dielectric, thermal, and mechanical stresses. Similarly, [14, Sec. 4] states that there are four key facets to a TX's functionality: EM circuit, current carrying circuit, dielectric withstand, mechanical withstand. The primary functional subsystems of a TX are listed in [11, Sec. 4] as the: *OLTC*, *Bushings*, *Cooling System*, *Oil Containment and Preservation System*, and *Active Part*. The latter consists of the dielectric system, mechanical structure, EM circuits, and current carrying circuits.

The *Bushings* and *Tap-Changers* (OLTC / DETC) are often considered as separate functional subsystems. They can have their own *Oil Containment and Preservation Systems* as well as be detachable from the TX *Tank*. For example, [15] calculates the health of the TX and the *Tap-Changer* independently before combining them based on their justification that there is a degree of independence between the health of these components. This thesis does not consider the *Bushings* nor *Tap-Changers*.

The oil insulation interlinks the *Active Part*, *Oil Containment and Preservation System*, and *Cooling System*. Its required properties for enabling the intended functionality of the *Active Part* are dependent on its integrity / purity, which is in turn is dependent on the other two subsystems. A faulty *Cooling System* can cause overheating which can accelerate the deterioration of the paper insulation, potentially leading to an accelerated failure of the *Dielectric System* for example. It is therefore important to appreciate that these subsystems are interacting in complex ways. A *Functional Transformer Model* to guide *Condition Monitoring Techniques* that "describes how the

transformer is broken down into subsystems, and defines the functionalities and failure modes of each subsystem" is conceptualised in [11, Sec. 3]. Fig. 2-3, redrawn from [11, Fig. 3.6], provides an illustrative example of their concept. One potential issue in creating a comprehensive version might be the highly interlinked nature of TX FMs. It is stated in [11, Sec. 3] that there is no singular perfect implementation of a *Functional Transformer Model* as it depends on the intended purpose. Factors such as the size of the fleet influences the level of detail needed in such a model.



Source: Redrawn from [11, Fig. 3.6], original from CIGRE © 2015

Fig. 2-3: Functional Transformer Model

2.1.3. Failure Modes and Stressors

Failure Modes Context

If a *Failure Mode* (FM) is a means by which an asset can no longer perform satisfactorily, it represents a binary threshold differentiating pre-failure and post-failure. However, there may be an underlying form of degradation occurring to trigger the onset of a FM. For example, contacts may gradually corrode or deteriorate until there is a FM of electric continuity of the circuit. Conversely, a falling tree branch may break the circuit continuity; this would have no precursors and is more "brittle" in nature. Four examples of evolutionary patterns of functional failures are given in [16, Sec. 1]: *Intermittent Failure Pattern*, *Binary Pattern*, *Fast Wear Pattern*, and *Slow Wear Pattern*. The timeframes are relative to observation resolution; as the sampling rate is increased, the more likely a failure pattern can be distinguished from the *Binary Pattern*. Therefore, by identifying precursors to FMs and monitoring them at appropriate temporal resolutions, preventative measures can be taken to either avoid or reduce their impact.

It is also possible to categorise by the *Failure Type*, these are listed in [17, Sec. 2.2] as: *Time-Based Failure*, *Utilisation Failure*, *Random Failure*, *Hidden Failure*, and *Asset*

Specific Failure. Understanding the nature of the *Failure Type* can enable the better selection of *Condition Monitoring Techniques*. *Condition-Based Failures* can be tracked via degradation monitoring for early warning whereas *Non-Condition-Based Failures*, such as lightning strikes, cannot be predicted and fall outside the remit of *Condition Monitoring*. However, there is a nuance that perhaps a degraded TX is more susceptible to failing under these random events. This is sometimes captured by recording within a TX's history the number of through faults it has experienced for example, partly as a marker of its potentially increased vulnerability [7, Sec. 4], [8, Sec. 3]. A common method to account for these kinds of events is using a probability of it occurring per time interval. This is essentially a function for *Probability of Failure* (PoF) accounting solely for time whereas *Condition-Based Failures* could instead have a function for PoF also accounting for the current *State of Health* (SoH). When aggregated, this can provide an indication of 'expected', unexpected failures. This insight can be used to pre-allocate contingency resources to then apply reactively to reduce the *Consequence of Failure* (CoF) rather than the PoF.

Failure Modes and *Failure Types* can broadly be attributed to *Failure Causes* or *Stressors*. It is important to distinguish these; as per [12, Sec. 5], "understanding how the failure occurs is useful in order to identifying the best way to reduce the likelihood of failure or its consequences". These are typically contributing factors to an asset's eventual failure, often accelerating its occurrence, though sometimes they may also be the direct cause of said failure. Different terms are used to represent a similar concept to *Failure Causes* or *Stressors* in literature. For example, [10], [13], [14], and [18] use the terms *Component Stressors*, *Influences on Rate of Failure Mode Progression*, *Aging Factors*, and *Stressors* and *Aging Mechanisms*, respectively.

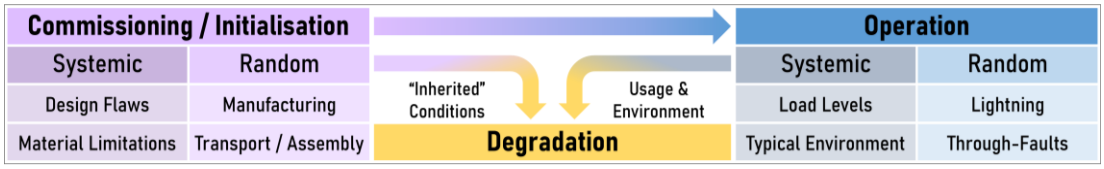


Fig. 2-4: Sources of Degradation

These *Failure Causes* or *Stressors* can be broadly split into *Systemic* and *Random* categories, where *Systemic* signifies those present in all assets with the same circumstances whereas a *Random Cause* or *Stressor* will be present sporadically. The eventual FM of an asset may be due to different *Failure Causes* or *Stressors* having

occurred at different points in its lifespan. Therefore, it can be useful to divide the asset's life into different stages. Fig. 2-4 presents a simple division, separating before and after *Commissioning*. However, some literature will include further divisions, for example Appendix 3 of [8] lists: *Transportation and Storage*, *Erection and Adjustments*, *Commissioning*, *Operation*, and *Maintenance and Dismantling*. Fig. 2-4 shows certain conditions are 'inherited' and cause a predisposition to given failures or degradation tendencies. An example of a *Systemic* cause at this stage may be a design flaw with a particular asset design, whereas a *Random* cause may be a mistake during manufacturing or damage caused by transport. During the *Operation*, the loading intensity can be considered a *Systemic* cause whereas a lightning strike is a *Random* cause in that it is generally unpredictable.

Although factors prior to *Commissioning* are highly relevant, they are often difficult to influence as they occur only once at the start of an asset's life in contrast to the factors occurring during *Operation*. For TXs, [9, Sec. 9] has further details regarding the different types of testing. It uses three categories: "factory testing when the transformer is new or has been refurbished, acceptance testing upon delivery, and field testing for maintenance and diagnostic purposes". These distinctions are relevant for better understanding and thus planning *Condition Monitorin*g.

Transformer Failure Modes and Stressors

Primarily sourced from [14], the following overview highlights some common thematic patterns. Section 2.3 will further detail gas-generating issues relevant to DGA. Different literature will suggest different categories for typical FMs. For example, [13, Sec. 7] uses the four categories: *Dielectric*, *Thermal*, *Mechanical*, and *External*. For comparison, [14] also includes *Physical Chemistry* and *Electrical* whilst excluding *External* whereas [19] adds *Unknown* to the list used by [14]. The organisation of categories is often driven by the principles of operation of an asset; assuming it fulfils its objective of ensuring all FMs are considered, the specific choice of taxonomy is inconsequential. This thesis uses the following categories: *Chemical* (*Contamination*), *Dielectric*, *Electrical*, *Magnetic*, *Mechanical*, and *Thermal*.

A functional failure occurs when a TX's operating stress exceeds its withstand strength for a given FM [13, Sec. 3]. Said withstand strength naturally decreases over a TX's life.

In addition, there are *Stressors* that can influence a TX's lifespan. Some highlighted in [18, Sec. 4] are: temperature, voltage, mechanical and electrical cycling of auxiliary components, non-seismic vibration, radiation, and environment (humidity, dirt, dust, and contamination). Some aspects not mentioned include loading, harmonics, through-currents, and voltage spikes. For example, [1, Sec. 1] highlights how TXs used for wind turbines have elevated gas levels due to the widely fluctuating loads. Additional influences factors listed in [13, Sec. 7] include design, maintenance, and protection. For example, [17, Sec. 1] highlights how some TXs have inherent design weaknesses, and that thermal ageing of paper is very design dependent.

Overview of Transformer Component Failure Modes

Table 2-1 relates FMs to specific TX components whilst indicating the likelihood of the relationship. It was modified from Table 2 from [13, Sec. 7] to show only the *Active Part* components. One limitation of Table 2-1 is that it does not include any distinction between *Defects* and *Faults*: an influential factor in decision-making. Table 4-1 from [14, Sec. 4] tabulates some of the functional *Subsystems* and their constituent components against reversible *Defects* and non-reversible *Faults* and FMs. However, one limitation to Table 4-1 from [14, Sec. 4] is that it covers only the TX's *Active Part*. Care should be taken when cross-referencing literature as terminologies can sometimes conflict. For example, Appendix 1 of [14] uses the term *Defect* to mean an unusual sample warranting further investigation. If an abnormality was then found, the *Defect* would also classify as a *Fault*. If the *Fault* requires the TX's removal from service, then it would classify as a *Failure*. Although FMs and *Failure Causes* can be difficult to distinguish, they should not be conflated [20, Sec. 3]. For example, an *Overheating* FM can often be due to a *Dielectric* issue not preventing arcing.

Table 2-1: Typical Failure Modes of Key Components

Failure Modes	Failures Detected	Windings	Core	Connections (internal)	Insulation	Oil
Components						
Dielectric Faults	Insulation Deterioration	●	●	●	○	●
	Moisture Ingress Content	●	●		●	●
	Tap-changer Condition / Problem	●				
	Oil Quality Deterioration	●	●		●	●
	Arcing / Electrical Discharge	●	●	●	●	●
	Connection Problem	○		●		●
Thermal Faults	Overheating / Auxiliary Cooling System Problem	●	●		●	●
	Low Oil Level	●	●		●	●
	Oil Circulation System Problem	●	○		○	○
Mechanical Faults	Winding Distortion		●		●	●
	Winding Looseness		●	●	●	●
	Core Looseness		●		●	
	Oil Leak	●	●			○
	External Damage / Disturbance		○	●	○	
External Faults	e.g., Animals				○	
	Through Fault, e.g., lightning strike or short-circuit	○	●	●	●	●
	Supply Faults, e.g., excessive harmonics / over fluxing	○	●	●	●	●
Legend:	● = Likely Relationship	○ = Less Likely Relationship				

Source: Modified from Table 2 from [13, Sec. 7], original from BSI © 2018

Transformer Subsystem Failure Modes

A brief overview of some of the common FMs in TXs is provided using TX *Subsystems*. More detailed information can be found in [14, Sec. 4]; the main source summarised here. Note these depend on specific asset design which often vary significantly.

Current Carrying Circuit

The *Current Carrying Circuit* can fail due to poor contacts causing excess heating, which can cause oil overheating and gassing, and increases in contact resistance and irreversible degradation of the contacts [14, Sec. 4]. This causes coking, oil breakdown, and potentially either open-circuit or short-circuit occurrences [14, Sec. 4].

Electromagnetic Circuit

The *Electromagnetic Circuit Defects / Faults* are typically attributed to either general overheating due to issues with cooling, or local *Core* overheating due to intended main magnetic flux, or unintended stray flux. Local overheating due to excessive eddy current losses can result in insulation deterioration, or the generation of gas, carbon,

and other degradation products [14, Sec. 4]. Closed loops created by stray flux, if accompanied with poor contacts, can result in overheating, sparking / arcing, and in insulation deterioration [14, Sec. 4]. A floating potential, due to ungrounded magnetic shields for example, can also cause sparking.

Mechanical System

The magnetomotive forces can also cause a FM via the *Mechanical System*. Large forces can be exerted onto the *Windings* and *Core* by through-faults which can cause deformations, damage to the solid insulation, or excessive noise or vibrations [13, Sec. 7]. Deformations can cause *Partial Discharge* (**PD**), or excess losses and heat generation [14, Sec. 4]. Additionally, a switching surge over deformed *Windings* can also lead to flashover between the *Coils* and excess gas evolution [14, Sec. 4]. Any movement could also affect the integrity of the contacts or solid insulation potentially causing arcing or open / short circuits.

Dielectric System

The *Dielectric System* primarily consists of *Solid Insulation* and *Liquid Insulation*. The *Solid Insulation* degradation is largely considered irreversible, and therefore, the maintenance of the *Liquid Insulation* is often undertaken to minimise the rate of *Solid Insulation* degradation [7], [8], [9], [11], [17]. As per [9, Sec. 6], "the life of the transformer is the life of the paper…". The *Solid Insulation* is typically paper and pressboard or more generally, cellulosic materials. As the *Solid Insulation* ages, it grows more brittle and prone to mechanical failure, as well as reduces in dielectric withstand capabilities [8], [13], [14], [21]. The *Liquid Insulation* similarly loses its dielectric withstand capabilities as it degrades and can become more acidic [8, Sec. 4], [14, Sec. 6]. Temperature, water, and oxygen accelerate degradation [14]. These drive pyrolysis, hydrolysis, and oxidation, respectively. Although in the case of hydrolysis, acids also play a significant enabling role [14, Sec. 3], [17, Sec. 1]. Particles, even non-conducting ones such as dirt, can decrease the dielectric strength of the *Liquid Insulation* and are therefore important to control within a TX [7, Sec. 12]. These particles can originate from "the components within the transformer itself, from arcing, fault degradation products in the equipment, or ingress during maintenance or repair" [8, Sec. 5]. Five forms that water can exist in a TX are listed in [9, Sec. 6] including as ice, free water, humidity, and dissolved in the oil.

Reference [8, Sec. 5] characterises the ageing and thus degradation of oil insulation as primarily due to oxidation, contamination, overheating, and arcing or discharge. As per [1, Sec. 4], the "characteristics of deterioration include sludge accumulation, weakened strength of insulation materials…, and shrinkage of materials that provide mechanical support". The materialisation of sludge is also dependent on temperature levels, with lower temperatures more readily forming sludge [8, Sec. 5]. A darkening of the oil, and in some cases suspended particulates, can be visible precursors to sludge generation [8, Sec. 5]. Two critical stages of dielectric withstand strength degradation are highlighted in [14, Sec. 4]: a *Defective* condition, and a *Faulty* condition. The former is characterised by non-destructive **PD** at operating voltage and a reduction in impulse withstand strength whereas the latter is characterised by destructive **PD**, as well as progressing surface and creeping discharges.

<u>System Interactions</u>

Often, different degradation pathways have accelerating feedback mechanisms [13, Sec. 7]. Fig. 2-5 shows a simplified selection of interactions between the systems. Examples of sources of stress on the dielectric, mechanical, chemical, and thermal systems are given, and then each system's interactions are listed. For simplicity, the electrical system is excluded, and it is assuming there are no relevant thermal–mechanical interactions. An example interaction is how overheating can cause gassing, which greatly reduces the dielectric withstand of the oil, potentially leading to further overheating and even a catastrophic thermal runaway [21, Sec. 4]. Another example is how the cellulosic degradation of the *Solid Insulation* can release water, which can decrease the dielectric withstand of the *Liquid Insulation*, the resulting increased temperature will then further accelerate the *Solid Insulation*'s degradation rate. Further examples can be found listed in [9, Sec. 6]. Models such as Fig. 2-5 are too simple to incorporate all relevant interactions. For example, there are many self-fuelling feedback mechanisms not listed such as heat generated due to a dielectric fault reducing the dielectric withstand of the *Liquid Insulation*, exacerbating the dielectric fault.

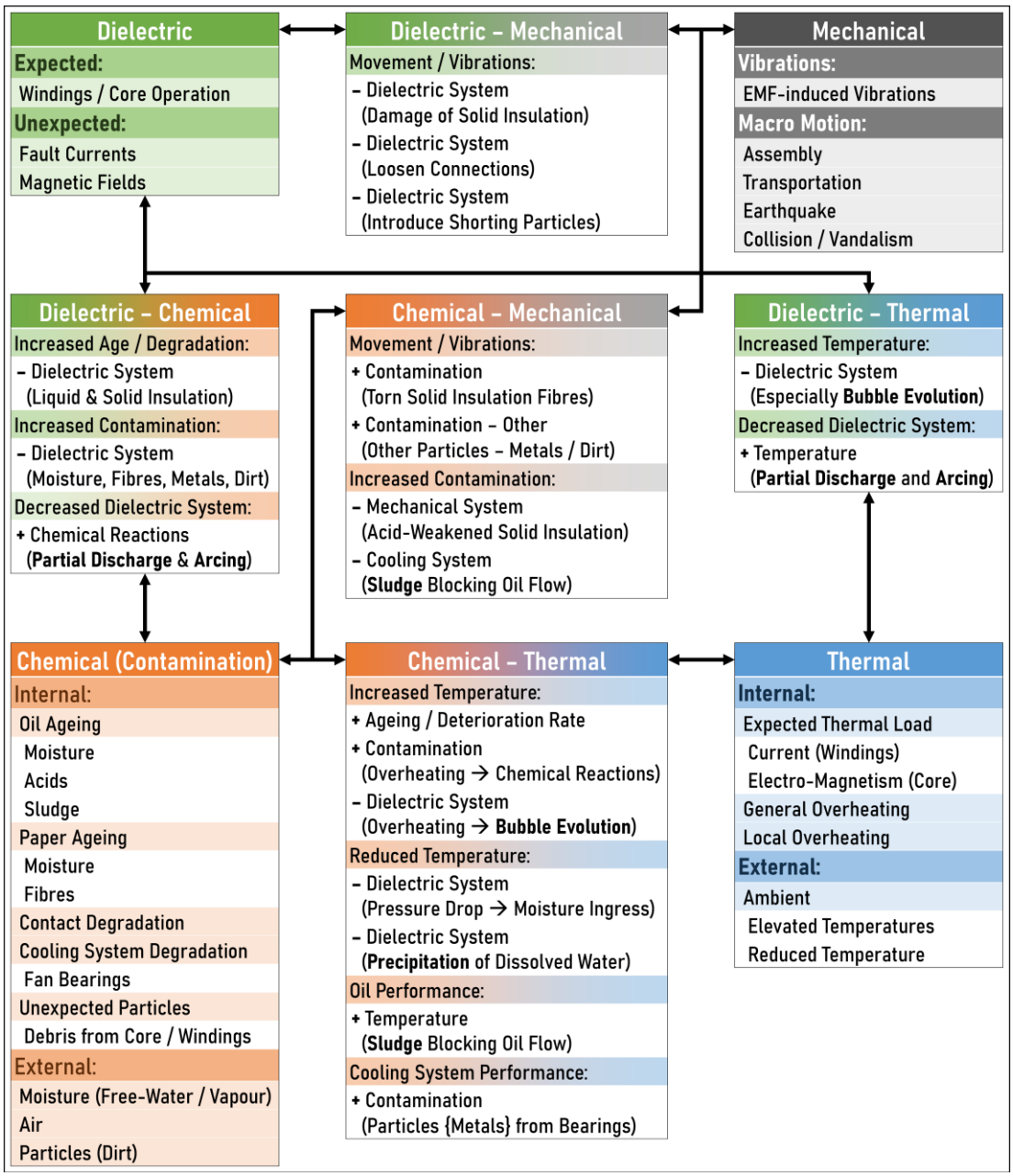


Fig. 2-5: Simplified Transformer Interactions Model

2.2. Condition Monitoring and Asset Management

2.2.1. Network Management Context

Grounding *Condition Monitoring* in the wider electrical network context is challenging. Four potential use-cases for *Condition Monitoring Interpretation* and *Diagnostics* are listed in [21, Sec. 6] as: condition assessment and life management; informing load planning; failure analysis; and preventing failures or unplanned outages. This broad scope results in many stakeholders for the outputs of *Condition Monitoring*. Table 2-2

provides an overview of the primary stakeholders in context of their specific interests based on a rewording of Table 2.1 from [11, p. 10]. *Degradation Evolution* was added to *Maintenance & Planning* as it is argued it can directly influence their decisions. The focus of this thesis is on the *System Operator* and *Maintenance & Planning*.

Table 2-2: Stakeholders of Condition Monitoring Outputs

Primary Stakeholder	Time Frame of Interest	Implications of Interest
System Operator Emergency Operation Emergency Maintenance System Operation	Immediate / Short Term	– Safety – Continuity – Reliability
Maintenance & Planning – Planned Maintenance – Replacement Planning – "Intensive Care" / "Early Warning"	Medium Term	– Maintenance – Degradation Evolution* – Short Term Replacement
Strategic Asset Management – Long Term Evolution – Grid Reinforcement / Extension – Replacement Strategy	Long Term	– Degradation Evolution – Maintenance Optimisation – Long Term Replacement

Note. Degradation Evolution was added to Maintenance & Planning, differing from source.

Source: Reworded from Table 2.1 from [11, p. 10], original from CIGRE © 2015

Another topic tangential to the thesis is [22, Sec. 6]'s *Condition Monitoring Systems*, relating to the practical implementation of the system. As [11, Sec. 2] emphasises, *Condition Monitoring* implementation is commonly inhibited by pragmatic system issues such as interoperability, infrastructure investment, and cyber security conflicts.

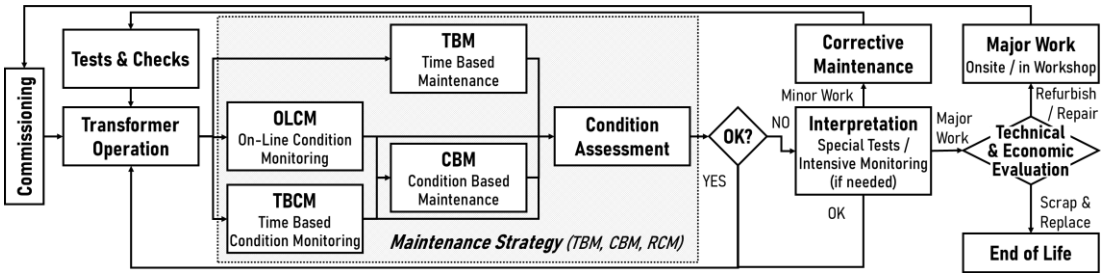
2.2.2. Condition Monitoring

Condition Monitoring in Asset Management

As per [23, Fig. 1], *Asset Management* is the "coordinated activity of an organisation to realise value from assets" where "realisation of value will normally involve a balancing of costs, risks, opportunities and performance benefits". The thesis focusses on *Asset Management* more in the context of ensuring expected performance to within an acceptable risk tolerance in a cost-effective manner. Other considerations, such as upgrading and capital investment management, are important but out of scope. The *Maintenance Strategy* is primarily responsible for this aspect of *Asset Management*, and therefore, *Repair*, *Refurbish*, and *Replace* are the main interventions of interest. These are supplemented by the option to increase *Condition Monitoring* by either increasing the frequency of existing *Condition Monitoring Techniques* or introducing additional monitoring techniques. Generally, there is a budget of money and personnel to allocate

to ensure the continued business-as-usual status quo to within an acceptable risk tolerance as defined either by the organisation or regulations. Therefore, the most needful of assets are to be prioritised in a cost-effective manner. Another aspect to consider is having the opportunity to schedule an intervention to reduce disruption and increase efficiency. This thesis considers *Asset Management* to be the process of this budgeting, planning, and prioritisation. However, fundamental to these are the topics of *Uncertainty* and *Risk Management*.

A major challenge of *Asset Management* is to cost-effectively account for *Uncertainty* as many critical parameters needed for optimum decision-making are either unknown or unknowable. Instead, estimates based on assumptions are used in their place. One of the primary purposes of *Condition Monitoring* is to inform *Asset Management* decisions by reducing *Uncertainty* of a given condition-related parameter. Without knowledge of an asset's condition, only the simplest protocols can be followed such as *Time-Based Maintenance*, replacement before *End of Life* (EoL), or repair after failure [8, Sec. 2], [14, Sec. 4], [16, Sec. 2]. These can be termed as *Maintenance Strategies*, and according to [8, Sec. 1], this involves any potential *Condition Monitoring* as well as *Condition Assessment*, as shown in Fig. 2-6 redrawn from [8, Fig. 1].



Source: Redrawn from [8, Fig. 1], original from CIGRE © 2011

Fig. 2-6: Transformer Operation and Maintenance Cycle

There are three *Maintenance Strategies* listed in [8, Sec. 2]: *Time-Based Maintenance* (TBM), *Condition-Based Maintenance* (CBM), and *Reliability Centred Maintenance* (RCM). In contrast, [16, Sec. 3] presents five categories, splitting RCM into three to also include *Risk-Based Maintenance* (RBM) and *Performance Focused Maintenance* (PFM). These strategies increase in complexity and cost-saving potential [8, Sec. 2]. They aim to establish the appropriate level of maintenance, the maintenance tasks, and when to perform them [8, Sec. 2].

TBM uses predetermined time intervals to justify interventions and includes actions to improve the asset's condition such as changing the oil for a TX [8, Sec. 2], [16, Sec. 2]. Two advantages to this approach are simplifying the planning process as well as the 'peace of mind' of being able to adhere to OEM's recommendations, which are traditionally based on regular interval-based maintenance [8, Sec. 2]. However, this approach is often deemed less cost-effective as the intervals are inherently set to be quite conservative relative to the actual expected lifespan to avoid too many failures [8, Sec. 2]. Depending on the application, this approach is also sometimes termed as *Frequency-Based Maintenance* (FBM) [16, Sec. 2]. By introducing *Condition Monitoring*, either in the form of *Time-Based Condition Monitoring* (TBCM) or continuous *On-Line Condition Monitoring* (OLCM), the generic intervention intervals can be adjusted based on more specific information regarding the given asset. This would include factors such as an asset's usage, occurrence of events, possible wear of moving or current switching parts or relevant applicable equivalent for a given asset, and the performance of similar equipment. Thus, the premise of CBM; maintenance carried out dependent on the equipment condition to reduce redundant interventions [8, Sec. 2], [16, Sec. 2]. However, it requires a more complicated planning process, and a method to measure and assess a metric(s) to determine the appropriate course of action [8, Sec. 2]. CBM is "often used within a time-based outage plan to defer maintenance to the next available outage", lessening the added complexity introduced to the planning process [8, Sec. 1].

Within CBM, although a sensor may provide data to suggest a potential developing issue, it does not inherently convey the significance of the issue regarding the reliability of the asset. RCM is focussed on identifying the most technically and economically effective maintenance actions to address this shortcoming of CBM [16, Sec. 2]. However, data suggesting imminent failure does not convey the asset's criticality to the organisation. RBM aims to account for these factors during the assessment of a given asset [16, Sec. 2]. RBM attempts to measure a metric more like *Risk*, which is typically considered as a product of the consequence of an event and its likelihood of occurring [8, Sec. 2], [16, Sec. 2], [23, Sec. 3], [24, Sec. 4]. For example, a TX deemed of high criticality might have a lower acceptable degradation level and thus a more onerous maintenance schedule. Therefore, like reducing the *Probability of Failure* (PoF), reducing the *Consequence of Failure* (CoF) is a viable way to lower *Risk*. This can be achieved via measures such as increasing redundancy or implementing automated load

shedding protocols [8, Sec. 3]. However, it is practically challenging to quantify all aspects of *Risk*, especially given factors such as the relative criticality of an asset, or costs associated with an intervention are specific to a given organisation [14, Sec. 7].

Performance Focused Maintenance (PFM) is also mentioned in [16, Sec. 2] as an even more comprehensive approach than RBM. It considers the organisational context in which RBM is to exist. As a demonstrative distinction between *Asset Management* in general and [8]'s RCM or [16]'s PFM, the latter two are focussed on optimising maintenance decisions whereas *Asset Management* may also include topics of maximising value of assets via other means. For example, it may be financially profitable to overwork the TX above its nameplate rating for a short duration despite its accelerated loss of expected *Remaining Useful Life* (RUL). This decision is unlikely prompted by RCM / PFM; rather, RCM would assess and account for the expected impact. Although, it could arguably be within the grander scope of PFM.

There are of course other frameworks for differentiating *Maintenance Strategies*. For example, [21, p. v] states that *Maintenance Technology*, here equated to *Maintenance Strategy*, has evolved through four levels:

- *Corrective*: Ensuring that equipment is operating and functional
- *Preventive*: Optimising the performance of the equipment
- *Predictive*: Diagnosing impending downtime for maintenance
- *Strategic/Optimisation*: Operational control and corporate-wide asset management.

Whilst plant optimisation systems are common, TXs remain challenging due to: "the reliability of the electronic equipment, cost of the monitors, continuing development of the sensors and monitoring systems, performance under harsh field conditions, lack of availability of field expertise, data collection, and interpretation" [21, p. v]. Additional context for a practical implementation in industry is given in [8, Sec. 3], which it defines as a generic *Maintenance Process*. It details how said *Maintenance Process* can be optimised and planned considering work coordination, staff availability, budgetary allowances and operational requirements or other constraints. This is a complex task, requiring multiple iterations to optimise, and is not generalisable.

Condition Monitoring Techniques

As per [13, Sec. 5], "the main objective of condition monitoring is to know about the condition of equipment, to be forewarned of possible failures, and to be able to carry out appropriate maintenance tasks at the appropriate time, i.e., condition-based maintenance". Five methods are listed in [17, Fig. 3] for what it terms as *Failure Mode Detection Techniques*. The thesis considers these equivalent to *Condition Monitoring Techniques* in this context. The five methods listed are:

- Periodic Inspection
- Periodic Operation
- Alarm / Indication / Metering
- Sample Monitoring
- Continuous Monitoring

These options are often most effective when used in combination. For example, *Periodic Operation* can ensure the asset can still operate as intended. This is most applicable to assets that may go unused for long periods of time, such as Circuit-Breakers (CB). Even *Continuous Monitoring* can be insufficient as it only captures data once the CB is activated—too late to pre-emptively accommodate. Further, comprehensive coverage via *Continuous Monitoring* can be cost-prohibitive. Therefore, options like *Periodic Inspection* can be invaluable additions to spot developing degradation in areas not well covered by *Continuous Monitoring*. The asset's relative value to operations can determine the techniques deployed. Even for high-cost TXs, [21, p. v] stresses that *Periodic Off-Line Diagnostic Testing* "still play an important role in industry".

As per [21, Sec. 3], *Intrusive* techniques will require "opening and / or exposing the interior of a transformer or its components" whereas *Non-Intrusive* techniques can be done without exposing the interior. This can be impractical for vacuum-sealed or gas-insulated assets. More generally, another drawback is that sometimes issues are introduced into the system via the inspection process, especially for *Intrusive* techniques that require disassembly and reassembly. This may result from incorrect reassembly, damage caused by the process, or the infiltration of contaminants.

For *Condition Monitoring* within a *Maintenance Strategy*, [8, Sec. 2] differentiates between *Off-Line* and *On-Line Condition Monitoring* depending on whether the TX can

remain operational during the monitoring. *Off-Line Condition Monitoring* is typically *Time-Based Condition Monitoring* (TBCM) with *Condition-Based Condition Monitoring* (CBCM) used for further clarification of the current condition. For example, a time-based gas sample tested in a laboratory may prompt further investigation. This also applies to *On-Line Condition Monitoring*, where a time-based visual inspection may prompt further investigation. As per [8, Sec. 2], *Continuous On-Line Condition Monitoring* is a sub-category of OLCM where an *Intelligent Electronic Device* (IED) provides measurement and control functions at a relatively high temporal resolution.

There is no specific minimum sampling frequency required to be considered *Continuous On-Line Condition Monitoring*. In the context of *Dissolved Gas Analysis* (DGA) in TXs, where lab-based sampling rate is typically measured in months, an on-line system's sampling rate of daily or more frequent is often considered tantamount to "continuous". When defining *Continuous Monitoring* specifically for TXs, [1, Sec. 4] states that it is at "very short intervals (e.g., daily or several time per day) ..." and that it could use "suitably connected on-line monitoring devices", making a distinction between *Continuous Monitoring* and *Continuous On-Line Condition Monitoring*. For reference, the survey of *Systematic Preventive Maintenance* for GSU and *Transmission* TXs in Appendix 1 of [8] saw a median DGA sampling interval of 1 year.

However, there remains some *Faults* out of scope that occur too rapidly. In these situations, other safety devices that are more reactive in nature, such as the Buchholz relay, would come into effect, followed by *Corrective Maintenance* if appropriate. For other assets, such as CBs, the most relevant measurements can only be taken when the CB is activated. In this case, "continuous" would be more in relation to capturing every activation and relaying the data in a timely manner back to an accessible source for future reference or automated processing. Occasionally, *On-Line Condition Monitoring* is meant as *Continuous* implicitly. For example, [21, p. v] equates the two when outlining the benefits of *Continuous On-Line Condition Monitoring*.

A benefit of *Condition Monitoring* highlighted by [22, Sec. 11] is the potential in identifying and distinguishing between reversible *Defects* and irreversible *Faults*, allowing for better decision-making regarding the best intervention. [21, Fig. 6] illustrates the benefits of *Condition Monitoring* in increasing detectable *Faults* and thus

avoiding an increased subset of catastrophic failures. However, it is important to note that *Condition Monitoring* cannot mitigate all risk. As per [21, Sec. 7]:
"It is unrealistic to expect a detection efficiency of 100%. Some faults can go undetected or develop at a rate too fast to allow for proper alarming and orderly removal from service.... The faults not detected include those that are instantaneous by nature, for instance an insulation breakdown following a lightning surge or severe short-circuit. Moreover, some components such as bushing shields are prone to sporadic failures that occur without any warning".

Lastly, aside from cost, another factor is the concept of erroneous interventions due to added screening (false positives)—commonly considered when assessing the benefits of medical screening amongst populations [25]. This is particularly relevant where the *Condition Monitoring* is *Intrusive* as this can increase downtime and the risk of introducing issues during the process. As per [18, Sec. 9], "many times, more damage is done by opening a transformer and doing an internal inspection than what is gained".

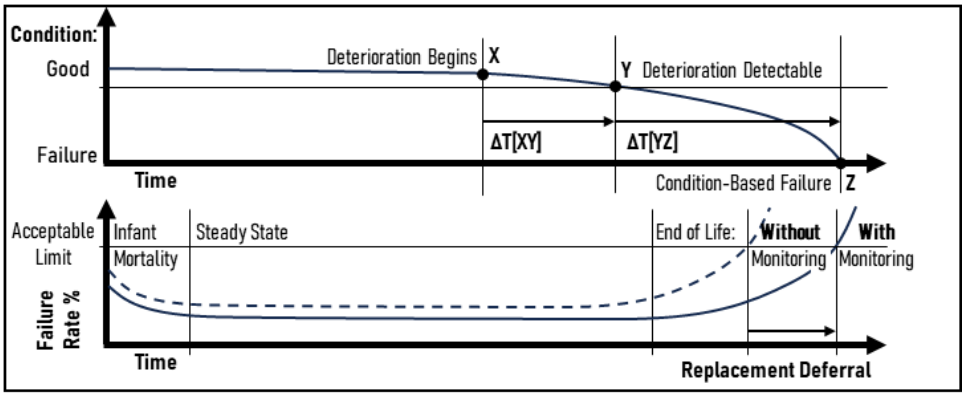
Condition Monitoring Feasibility

There are two prerequisites for successful *Condition Monitoring*: being technically feasible, and being economically justifiable [8, Sec. 2]. In this context, it is assumed that all the aspects of CoF have already been converted into a quantifiable economic cost, including safety and environmental considerations. An approach such as a *Cost-Benefit Analysis* (CBA) would determine whether the expected cost of inaction exceeds that of action to justify an intervention. *Technical Feasibility* is more related to the process of identifying the precursors to a given FM in time to act. [8, Sec. 2] suggests three criteria for *Technical Feasibility* with reference to [8, Fig. 2], redrawn in the top plot of Fig. 2-7, that is paraphrased to:

- *Sensitivity* – detectable condition change should be relatively small compared to the condition change required to cause a failure.
- *Forewarning* – change detected should allow time for preventive action.
- *Frequency* – change detected should be measurable at time intervals shorter than that required for the condition to deteriorate from "Good" to "Failure".

Knowledge to interpret the data is another requirement listed in [11, Sec. 2]. According to [13, Sec. 10], the point **Y** in Fig. 2-7 would be known at the *Point of Potential Failure*,

P, and point **Z** would be *Point of Functional Failure*, **F**, with the gap between them being termed the **P–F** interval. It is recommended in [13, Sec. 10] that the sampling frequency should be such that at least two successive measurements are obtainable within the **P–F** interval, ideally three to five to better cope with noisy signals. Increasing the sampling frequency makes it more likely to meet the *Technical Feasibility* criteria. Although whether the detectable condition change is sensitive enough is often more related to the nature of the FM and the available technology relating to *Condition Monitoring* of its precursive symptoms. In some cases, *Continuous OLCM* technology may not yet be sufficiently developed, at least at a cost-effective price point, to meet point one whereas other methods, such as laboratory testing, may have the required sensitivity to detect deterioration at an earlier stage.



Source: Redrawn from top: [8, Fig. 2], original from CIGRE © 2011, bottom: [21, Fig. 7]

Fig. 2-7: Theoretical Asset Degradation and Effect of Monitoring on Transformer Life

Continuous OLCM's greatly reduced sampling intervals when compared to manual sampling therefore have many benefits as outlined in [8, Sec. 2], [21, Sec. 4]. For instance, it can potentially identify incipient *Faults* and provide early warnings if problematic trends develop. Should a failure occur, the high-resolution data that was automatically stored can provide insight as to the immediate precursors for future learnings. Furthermore, the resources previously dedicated to manually monitoring the unit can now be diverted and an occasionally relevant consequence as per [1, Sec. 4] is that the risk associated with in-person *Condition Monitoring* of a potentially dangerous (suspected of catastrophic failure) asset is removed. Lastly, *Continuous OLCM* can lower the perceived risk levels for the lifespan of the asset. This can allow for not only the deferment of asset replacement, but also potentially facilitate higher load levels if the risk is deemed manageable given the extra monitoring, as shown in the bottom plot of Fig. 2-7 redrawn from [21, Fig. 7].

2.2.3. Condition Assessment

Condition Assessment Relative to Asset Life

Different data sources may be assessed to form a *Condition Assessment* representing an asset's condition for the pre-defined timeframe. It is stated in [7, Sec. 1] that the "essence of condition assessment is to identify the indications that can be used to determine (and quantify where possible) the extent of the degradation…". The asset's condition should be considered in the context of its *Operating Conditions* and *Asset Management Information* to inform realistic expectations for the asset.

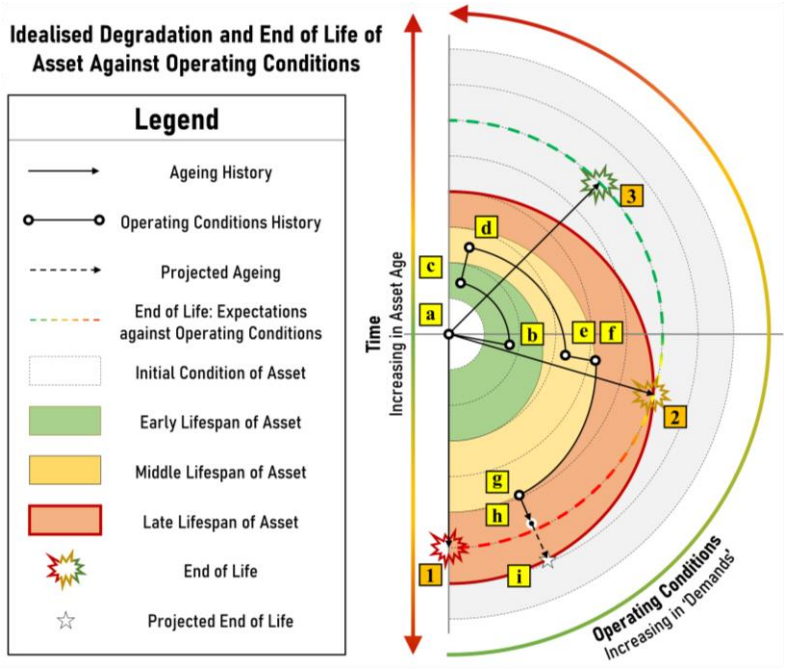


Fig. 2-8: Idealised Asset Degradation and End of Life Against Operating Conditions

Fig. 2-8 maps a simplified model of possible degradations for an asset. The white innermost semicircle represents the starting condition of the asset at *Commissioning*. The dotted semicircles represent equidistant time intervals of an asset's *Age*, similar to a contour map. The green, yellow, and red shaded regions then represent arbitrary thresholds of degradation. The outermost dark red boundary represents the EoL. The asymmetry is due to differences in the metric labelled *Operating Conditions*, where anticlockwise rotation represents more demanding conditions. The grey shaded area represents the loss of potential life compared to ideal circumstances due to these demands. These idealised thresholds assume the *Operating Conditions* metric fully explain variation in EoL expectations. The *Operating Conditions* can include factors such as working at a higher load-level or working in unideal environments such as near

the sea. These would include influences labelled as *Operation* in Fig. 2-4. In practice, the radii of the boundaries would vary. For example, the starting condition, termed *Commissioning / Initialisation* in Fig. 2-4, can vary depending on factors such as the TX manufacturer, damage incurred during transport or installation, or material defects.

The point labelled **1**, represents an asset's EoL having operated under ideal conditions. Its life is represented as a black arrowed line from the origin. Given these conditions, this asset reached EoL earlier than expected, whereas under different conditions, this could have been a typical *Age* (point **2**) or even better than expected (point **3**). It is therefore important to consider how expectations are set and that samples within a population are relevant and representative of one another. Points **a–i** show an example asset life. Between points **a–b** the asset is operating typically before something prompts an increase in work intensity. For example, a failing nearby asset increases this asset's load. The ageing trajectory is altered from point **b** to **c** and continues to point **d** until the asset is returned to typical operating conditions. The asset experienced accelerated *Ageing* during this time. At point **f**, it may be decided the asset is nearing EoL and this should be avoided for now, so the asset is operating more lightly. This is represented by the change in trajectory from point **f** to **h**, where it continues to point **h**, which represents its current state. Based on this operating condition, it can be estimated for the asset to fail at point **i**. An asset's expected life can be forecasted with knowledge of its current condition and operating environment.

Fig. 2-8 uses *Time* and *Age* in a more abstract manner than their SI units. For example, moving to a harsher condition seemingly reverses *Time*; this is because similar *Ageing* could have occurred in less equivalent *Time* under said conditions. It is more akin to *Apparent Age* or *Effective Age* as defined in [7], "the condition assessment information in a corrected age that may be used to estimate the present and future probability of failure based on statistical analyses".

Condition Assessment Characterisations

There are many similar, overlapping terms that are occasionally used interchangeably in the literature which can cause confusion. The goals of *Condition Assessment* are based on the *Condition Assessment Scope* and fulfilled via a *Condition Assessment Methodology*. This *Condition Assessment Methodology* will select suitable *Condition*

Assessment Techniques to interpret data collected via suitable *Condition Monitoring Techniques*. The relevance of the latter distinction is that, for example, often more data is collected via *Condition Monitoring* than is actively used for *Condition Assessment*. This may have been due to pre-emptive 'futureproofing' measures, where it was envisaged that the additional data may eventually be incorporated into the *Condition Assessment*, or simply the volume of data is too large to process or more precise than needed. Nevertheless, the more general topics of *Condition Assessment Techniques* and *Condition Monitoring Techniques* are similar, as the latter can limit the former.

Different available *Condition Assessment Techniques* are outlined in Table 1–1 of [7], categorised based on their ability to detect *Faults*: *Logical Reasoning*, *Inspections*, and *Basic* and *Advanced Condition Assessment Techniques*. It also highlights that not all *Condition Assessment Techniques* are of equal value. Often, the most informative techniques are the most expensive, and so are applied only when there is a cause for concern. During the initial construction and *Factory Testing* are the other times that more advanced and/or *Intrusive Condition Assessment Techniques* are applied as it is a controlled environment with access to the necessary equipment. More information specific to *Factory Testing* for TXs is found in [10, Sec. 6.2] and [14, Sec. 5.2].

Logical Reasoning is relied upon to infer the condition if it is not monitored. *Inspections* are a form of *Condition Monitoring*, though their subjective nature often requires some *Condition Assessment* to be performed by the inspector at the time. *Basic* and *Advanced Condition Assessment Techniques* can be any of the *Condition Monitoring Techniques*: *Alarms*, *Indicators*, *Metering*, *Sampling Monitoring*, *Continuous Monitoring*, and *Periodic Operation*. The distinction between *Basic* and *Advanced Condition Assessment Techniques* is unclear. Generally, the cost and time associated with *Basic Condition Assessment Techniques* should be less to offer utility. Although Table 1–1 of [7] conveys the general *Diagnostic* value of each technique, it does not elaborate on **how** the condition is assessed. It only provides a qualitative scale for the reliability and relevancy of the assessment. A primary limitation in *Condition Assessment* is the *Condition Monitoring Techniques* used and their relevancy, accuracy, reliability, frequency, and the domain knowledge for interpreting them. This closely resembles the *Condition Monitoring Feasibility* topic. These factors influence which *Condition Monitoring Methodologies* are the most appropriate. Another relevant consideration is the

relationship between the sensor data and the decision to be made. If a simple and direct relationship is present, there is no need for complicated analytic processes [11, Sec. 3].

The term *Interpretation Techniques* is used by [11, Sec. 4], where it divides them into: *Knowledge-Based Techniques*, and *Data-Driven Techniques*. It states that *Knowledge-Based Techniques* are aimed at encoding domain expertise and replicating their reasoning process. *Data-Driven Techniques* are instead aimed at encoding "lower-level pattern matching facets of intelligence". It also provides examples of *Knowledge-Based Techniques*: causal models, expert systems, and fuzzy logic, and examples of *Data-Driven Techniques*: neural networks, multivariate analysis, rule induction, and Bayesian networks. In theory, a *Data-Driven Technique* would have its conclusions repeatable given the same dataset whereas a *Knowledge-Based Technique* is relying on external information not necessarily present within the available data.

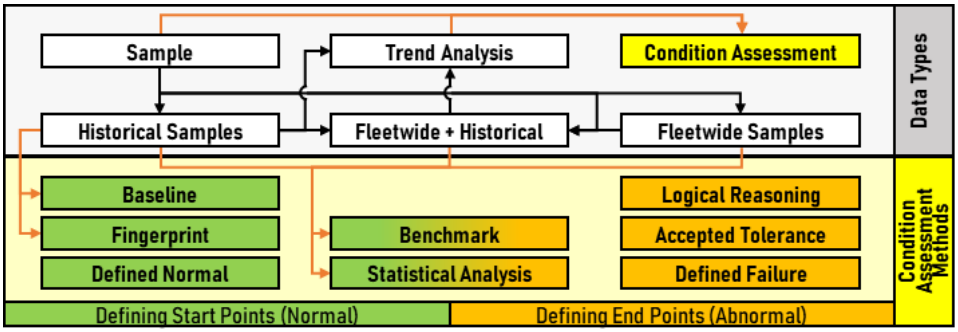


Fig. 2-9: Examples of Condition Assessment Methodologies

Fig. 2-9 diagrams some potential *Condition Assessment Methods*, based primarily from [14, Sec. 4.3] with some added details from [11, Sec. 4]. *Condition Assessment* is done with reference to either an assumed healthy status, or an assumed unacceptable status, or both. If a healthy parameter is known *a priori*, it can be used for comparison as a measure of degradation. For example, mineral oil may be sold to a specific standard with known tolerances. This is a form of *Knowledge-Based* analysis as there is not necessarily an initial data point available demonstrating these tolerances. Similarly, if the *Condition Monitoring Technique* relates tractably to a FM and its interpretation is known, then it may be possible to have thresholds signifying different states of degradation. This is termed as *Quantification of Defective Condition* by [14], and it would also constitute a form of *Knowledge-Based Technique*. In Fig. 2-9, it is termed *Defined Failure*. Paper degradation in TXs is an example of this, where the degree of polymerisation can be directly related to the paper health and thus have interpretive

thresholds [7], [8], [9], [10], [11], [13], [14]. Often, an *Accepted Tolerance* based on experience or warranties is relied upon instead. For example, a particular component may be rated for a certain number of operations and that can be used as a scale.

In practice, there are often naturally occurring variations between manufacturers or environments, behoving more specific thresholds quantifying degradation as not all the influencing factors indicate degradation. An example provided by [9, p. 117] is how the "Westinghouse 7M" series TXs' design causes more ethane and ethylene gas generation than some other designs. If this is a known issue of limited influence on the RUL of the asset, the thresholds for these gases for this specific design could then be loosened. *Fingerprinting* is a method for determining more applicable values, where comparison is made to a fault-free, as-new asset. In this context, assets would be compared to as relevant a 'fingerprint' result as possible. Fingerprint results are referred to by [14, Sec. 4] as reference results valid for all TXs of same design. This ideally requires the results of the *Diagnostic* test to be solely related to the characteristics and condition of the TX, and independent of the measurement system.

Sometimes, *Condition Monitoring Techniques* change over time or are added retrospectively to an asset fleet [8, Sec. 3]. This means that the original values from an as-new asset are no longer available. Instead, values could be compared to a specific instance where it is assumed that the asset is functioning as intended. This is called a *Baseline* or *Benchmark* approach by [14, Sec. 4]. The challenge would be knowing for certain that there are no *Faults*. Also, as an asset is used differently, they may deviate from their original starting point. This means it may be more likely that a *Baseline* or *Fingerprint* reference is less relevant to different assets. It is stated in [22, Sec. 9] that *Benchmarking* or *Fingerprinting* can improve accuracy and reliability of interpretation, and it highlights several *Condition Monitoring Techniques* for TXs that are particularly aided with the presence of such results.

Derived values from causal models may be used when observed values are not directly related to a FM [11, Sec. 4], [21, Sec. 5]. This is also a form of *Knowledge-Based Technique* as it requires domain knowledge. Desired unobserved values are predicted by leveraging a known relationship they have with other observed values. For example, in TXs, the hot-spot temperature within the oil is often predicted based on more readily

available parameters such as the top oil temperature, loading and the difference between the incoming and outgoing coolant temperatures [7], [8], [10], [11], [14], [21]. This relationship could also be discovered via a *Data-Driven Technique* assuming sufficient data, though determining causality then presents an additional challenge.

A more longitudinal approach can be taken by analysing how an asset progresses over time through *Trend Analysis* of historical samples. *Logical Reasoning* can be applied to *Trend Analysis* to infer if a parameter is changing at a rapid rate, especially if at an accelerating rate, it is likely a cause for further attention. This is echoed in [14, Sec. 4] which states that *Trend Analysis* does not necessarily help distinguish 'faulty' indications from 'unusual' results but that the "occurrence of a rising trend, particularly when the rate of change is increasing, is probably a definite indication of a serious problem or at least something to be investigated further" in the context of TXs. There is additionally still the need to establish a threshold to act and the requirement of sufficient sampling frequency to identify *Faults* in time.

Benchmarking can be used to establish (ab)normality. A sample is compared to a larger population. This population may consist of historical samples from the same or other relevant assets. Ideally, the population is as relevant as possible to the assessed sample. A very simple approach may be to select the worst N performing assets to investigate further. This works well when there are pre-allocated budgets. A variation would be to select based on a percentile-based threshold, "concentrating maintenance efforts on the … transformers most at risk" [2, Sec. 8]. The drawbacks are that by definition, assets will be flagged even if they would otherwise be considered healthy to meet the quota, similarly, if an excessive number of assets are unhealthy, they may not be flagged. For this reason, [2, Sec. 8] states that these are "preferably to be considered as initial guidelines…. They shall not be used to ascertain whether or not a fault exists…".

Though not a method explicitly listed in [14], it is possible to analyse both cross-sectionally and longitudinally. For this, an entire fleet (cross-section) is analysed over time (longitudinally). This could be indicative of the overall fleet's health. Analysed retroactively, if it appears to be degrading over time, it could indicate a need for greater maintenance / replacement investment. Analysed projecting into the future, it could enable more sophisticated prioritisation of assets and better planning opportunities.

Statistical Analysis is another method highlighted, although it is very broad in its scope and could arguably encompass *Benchmarking* as it can be a means of automatically establishing thresholds. The most common example is arguably the use of standard deviations of an assumed normal distribution representing the population to demarcate ranges indicating potential outliers based on the assumption that an asset should operate reasonably consistently over time. Statistical descriptors can also be used, such as variance or differences between consecutive samples in addition, or as an alternative, to absolute values. This overlaps with *Trend Analysis*. The alternatives to *Statistical Analysis* for establishing limits could be the *Knowledge-Based Quantification of Defective Condition*, and simple approaches such as selecting the top *N* values to be assessed based on resource availability. The latter would be a *Data-Driven* approach that is cross-sectional in nature as it is relies on comparing an asset to its peers.

Condition Assessment Scope

Generally, an asset can be expected to operate as intended for much longer than it is to operate in a faulty manner. For TX DGA, [1, Sec. 4] states: "the interpretation ... is based on the premise that a liquid immersed transformer in sound condition generates little or no fault gas under normal operating conditions". Broadly, *Condition Assessment* can have the goals of *Anomaly Detection*, *Fault Detection*, *Fault Identification / Diagnosis*, and even *Prognosis / Forecasting*. These can be considered ordered in increasing levels of utility to the decision-maker, assuming comparable accuracy and reliability. An ideal system would provide timely, actionable information to the correct decision-maker with relevant context and justification on demand.

Anomaly Detection

Based on the assumption that an asset is expected to perform in a reasonably consistent and typical manner when in a healthy state, *Anomaly Detection* looks for samples that seem *Anomalous* or more specifically in this context, *Atypical*. *Anomaly Detection* is also termed *Out-of-Distribution* (OOD) *Detection* in some contexts [26, Ch. 19] which has a more intuitive connection to the term *Atypical*. There is no presumption that *Anomaly Detection* highlights only cases deserving of intervention, i.e., there is no distinction made between benign and malign *Anomalies*. *Anomaly Detection* is described in [11, Sec. 4] as "the most basic type of analysis, where deviations away from the norm are identified (but not explained)". It is hoped that a subset of the *Anomalous*

cases are faulty assets, but this depends on the relevancy of the assessed metrics to *Faults*. The better the alignment between the metrics being measured and their reflection of *Faults*, the greater the overlap of the flagged *Anomalous* samples and faulty assets. For DGA in TXs in particular, [1, Sec. 4] states: "experience has shown that not all abnormal gassing events are necessarily related to transformer deterioration or permanent damage". An example is a noisy sensor: unusually high noise can see a sample flagged as *Anomalous* whereas an initial denoising step may have prevented said flagging. Here, the denoised metric is better aligned with the desired outputs.

Fault Detection

Fault Detection can be considered a form of classification [26, Ch. 1], where the goal is to highlight samples that show signs of a *Fault*. This is more discriminating than *Anomaly Detection* as it is intended to ignore benign *Anomalous* samples. A sample does not necessarily have to be *Anomalous* in the sense of being a statistical outlier compared to the population, but that the specific combination of values corresponds closely to a *Fault*. If the metrics are highly aligned with *Faults* and the population are largely in a non-faulty state, then *Anomaly Detection* will perform similarly. Metrics well-aligned with *Faults* can have very simple yet effective interpretations. For example, oil temperatures in a TX can very easily be linked to *Faults* related to excessive oil temperatures. Although, the cause is not necessarily as obvious.

Fault Identification / Diagnosis

Fault Identification or *Diagnosis* is a form of multinomial classification intended to categorise detected *Faults*. *Anomaly Detection*, in contrast, is more typically a binary classification. Following the example of the oil temperatures in a TX, a *Diagnosis* would attempt to link the elevated oil temperatures to a cause. This could be using domain knowledge, experience, or cross-examination with other data sources. For example, if there is no longer a current passing through the cooling fans and there are elevated oil temperatures, then the cause could be linked to the cooling fans. The terms *Diagnostic* and *Failure Cause Analysis* are defined in [21, Sec. 3] as an "interpretation of the data supplied by the monitoring system", and as "the diagnosis of failures or malfunctions" to "draw a conclusion as to the cause of the failure or malfunction and thus replace or supplement the troubleshooting phase of corrective maintenance", respectively.

It is stated in [21, Sec. 6] that "diagnostic methods should distinguish between changes that are 'noise', those of minor consequence, and those worth of immediate attention". This implies a measure of *Fault Severity* and overlaps somewhat with the next category of *Prognosis / Forecasting*. An implementation of this can be the relative importance of a *Fault Type*, or the extent to which a given *Fault Type* is present or has persisted. Following the example of the oil temperatures in a TX, this could be ranking the importance of a faulty fan against other potential issues such as overheating due to arcing, or attempting to quantify the extent the fan is faulty; perhaps by comparing expected and actual generated air flows. Creating a comprehensive scoring system that can order all these aspects can be challenging.

Prognosis / Forecasting

Prognosis is the evaluation of the data to predict the likely progression of the situation. This can be considered a type of *Forecasting* as a given metric is projected into the future. However, *Prognosis* does not necessarily require a granular projection of time-series data, it could instead simply be a lookup table mapping *Diagnoses* with failure rates. Following the example of the oil temperatures in a TX, a *Prognosis* could attempt to link the elevated oil temperatures to the estimated EoL. A *Diagnosis* could help refine the *Prognosis* as different causes would likely have different implications. For example, a faulty fan may simply cause an accelerated EoL whereas if the elevated oil temperatures were due to internal arcing, a potentially much more catastrophic and near-term EoL may occur. The use of a unifying metric such as EoL can help overcome the challenges in creating a scoring system.

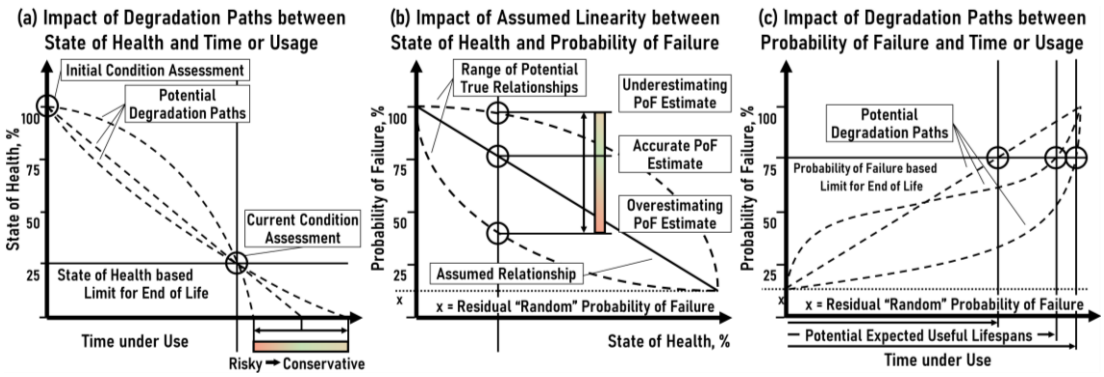
Prognosis is central to *Condition Assessment* as it is the basis of decision-making and closely tied to *Detailed Interpretation*. *Anomaly Detection*, *Diagnosis*, and *Prognosis* are summarised as "identifying there is a problem, recognising what the problem is, and predicting how much time remains in order to correct it" in [11, Sec. 4], respectively.

Condition Assessment Metrics and Indices

Condition Assessment Metrics

Condition Assessment requires a metric as an output represented on either qualitative or quantitative scales. In general, they represent *State of Health* (SoH) of the asset based on the captured *Condition Monitoring* data. SoH measures current degradation where

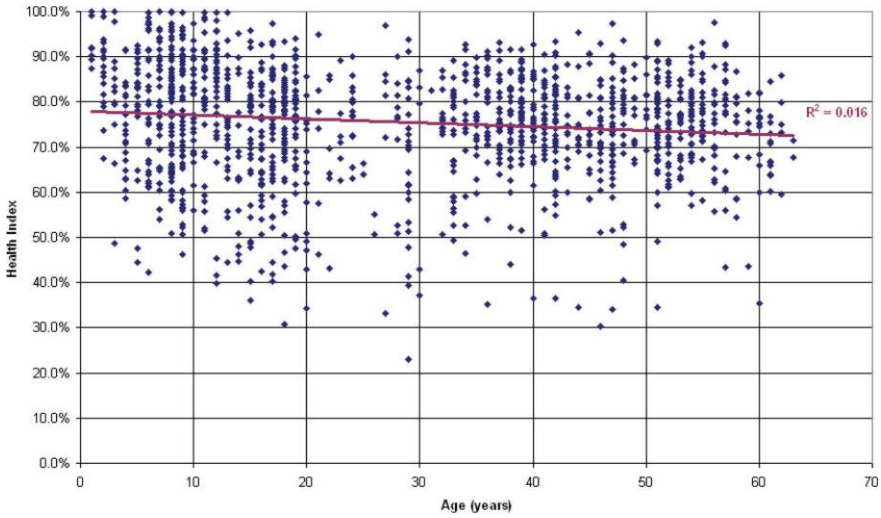
worsening conditions tend towards zero. It is not always clear how to quantitatively map measured conditions to such a scale, and linearity over time cannot be assumed. Even with constant operating conditions, some assets will not degrade linearly but rather tend to accelerate over time. Fig. 2-10(a) shows an abstract asset degrading over time. The degradation path can be one of many, as illustrated via the multiple dashed lines reducing in SoH as the *Time under Use* increases. Fig. 2-11 from [11, Fig. 4.13] provides an example of *Health Index* against *Age* of TXs and shows the expected downward trend, but also a large degree of variance between specific assets even of similar age.



(a): Impact of degradation paths between state of health and time or usage.
(b): Impact of assumed linearity between state of health and probability of failure.
(c): Impact of degradation paths between probability of failure and time or usage.

Fig. 2-10: Relationships between Health, Probability of Failure, and Time / Usage

Arguably, the main relevance of SoH is that it is assumed to be related to *Probability of Failure* (PoF). PoF is a measure of the likelihood that an asset will fail within a given time or usage interval, this accounts for the fact that assets in even seemingly good condition, or high SoH, may fail. PoF is expected to increase over time without intervention, though not necessarily linearly. Similarly, PoF and SoH often do not have a linear relationship, instead PoF tends to accelerate as SoH worsens. This is in part due to safety margins built into designs. Fig. 2-10(b) highlights the potential for misestimating PoF if assuming a linear relationship with SoH. This is shown in Fig. 2-10(c), where multiple degradation paths are plotted. Once again, the variance in expected useful life due to degradation paths is highlighted. Even at an 'as-new' SoH, PoF is rarely zero due to residual *Non-Condition-Based Failures* and random chance. Additionally, there is often an increased PoF during the initial stages due to inherent defects in assets. When viewing collated fleetwide failure rates, this would be represented with the "bathtub" concept of life expectancy [19, Fig. 7].



Source: From [11, Fig. 4.13] from CIGRE © 2015

Fig. 2-11: Example of Health Index Versus Age for Power Transformers

Condition Assessment Indices

For complex assets such as TXs, there are often numerous data sources referenced to determine its state. Consolidating various data inputs into a cohesive asset descriptor for actionable decisions can be challenging, especially when managing multiple assets [7], [11], [16]. This motivates simplifying metrics to condense the overall data volume. For example, it is common to produce a scoring index for ranking relative asset priority for a given intervention. The challenge depends on the complexity of FMs within the asset and the extent of monitoring. Compressing larger data volumes or more complex models increases the likelihood of losing or "masking" valuable information [7, Sec. 2]. This is exacerbated if many competing FMs can coexist in parallel.

It is challenging to consolidate *Condition Monitoring* data into what [7, Sec. 2] refers to as *Transformer Assessment Indices* (TAIs). There is clear value in assigning a simple score by which assets can be ranked and prioritised for *Asset Management*; however, it is easy to create a misleading metric if not taking due care. Two causes are the inadequate communication of the reliability or confidence of the outputted metric, and the fact that assets can have multiple FMs [7], [24]. A key recommendation of [7, Sec. 2] is make explicit the purpose and scope of a given metric. As examples of potential TAI purposes, [7, Sec. 2] lists the following *Indices*:

- *Reliability / Health* – likelihood of failure or expected time to failure,
- *Replacement* – those most benefiting from replacement,
- *Repair* – those most benefiting from repair/ non-essential maintenance,

- *Refurbishment* – those most benefiting from refurbishment,
- *Composite* – combination of the above, perhaps guided by financial considerations.

Here, *Refurbishment* differs to *Repair* by being preventative maintenance or repair. In contrast, [15, Sec. 4] instead treats refurbishment as an alternative to replacement.

Although each TAI targets a different action, they are related. For example, there may be no benefit to refurbishing a component within an asset if there is a different developing irreversible FM gating the expected EoL. A simple example elaborating further on this topic is provided in [17, Sec. 2.3.1]. When discussing EoL, it is primarily meant *End of Functional Life* as characterised by [27, Sec. 1]. However, [27, Sec. 1] highlights there are also other variants such as *End of Economic Life* and *End of Reliable Life* that may be relevant. The relevant timeframe for the TAI metric should also be made explicit. For example, an asset may be at present considered at a worse SoH to another, but that other asset may be degrading at a faster rate. It is therefore essential to understand the timeframe in which to comparatively assess these two assets to provide the expected relative indexing score.

TAIs can range from highly qualitative scales such as "good" to "bad", to quantitative scales such as 1–0. This is partially driven by the discussed practical challenges in quantifying SoH and PoF, so some TAIs instead use more abstract metrics [8, Sec. 3]. For example, a 1–0 scale or similar where the values are not necessarily quantitatively comparable, only qualitatively so. In other words, a lower value could be seen as worse, but not necessarily that a value half the other is twice as bad, [11, Fig. 3.8] is such an example. Another example of a dimensionless scale is from [1], which provides an output of one of three *DGA Status* levels, each with a list of potential recommended actions. These more abstract metrics are still useful for comparing and then prioritising assets within a fleet [7, Sec. 2], [8, Sec. 3].

There are also potentially other, more specific, outputs depending on the *Condition Assessment Scope*. For example, *Diagnostics* might include the *Diagnosis*, its confidence level, and the indicator of its severity on SoH / PoF / RUL. Similarly, *Anomaly Detection* could include the extent and confidence to which something is considered an outlier. These outputs provide additional context or insight where SoH / PoF / RUL cannot be accurately calculated.

It is the combination and complex interaction of these facets that make designing an effective TAI challenging. In practice, different *Condition Monitoring Techniques* vary in scope and capability for *Detection*, *Diagnosis*, or *Prognosis*. Furthermore, fleets may consist of TXs constructed decades apart [8, Sec. 3], meaning TXs can vary in designs and materials—affecting their expected lifespans.

2.2.4. Transformer Condition Monitoring Programmes

Developing a Condition Monitoring Programme

Balancing the economic and technical facets to select the appropriate *Maintenance Strategy* and *Condition Monitoring Techniques* for a given fleet of assets and even components within assets is clearly a complicated task. Therefore, it can be useful to take a systemised approach to ensure a developed *Condition Monitoring Programme* provides adequate coverage. Conceptually, the goal would be to consider the FMs, the symptoms that they may exhibit as they develop, and a means to monitor or measure said symptoms [7], [8], [11], [16], [17]. It is stated in Annex B of [7] that the steps involved in developing a TAI would be to first determine its purpose, identify the *Failure Modes and Mechanisms* within scope, determine how each FM would be assessed, and then design a calibrated system for categorising said FMs.

A dual-perspective methodology is provided in [11, Sec. 3]; looking "bottom up" then "top down". The more inductive perspective mentioned begins with the raw data from the sensor(s) and builds upon it with further analysis and interpretations to "arrive at physical properties, defects, failure modes, transformer status, associated risks, maintenance and replacement needs". The more deductive perspective begins with relevant stakeholder / decision maker drivers. Technology, budget, or risk were the given examples, although legislation is another that could have been included. It then states that from these needs, the required information and therefore data / analysis can be deduced. The actual decision-making process is stated to be outside of the *Condition Monitoring* process and is separated in their illustrative example of their described hierarchy [11, Fig. 3.3]. In addition, [17, Sec. 2] states that—with reference to [28, Sec. 5.2.4]— "rather than identifying every single possible cause for all failure modes, the level of detail should be reflective of the failure mode effects and their severity". The determining of the scope and appropriate level of detail for a given FM could be done

based on the analysis of failures, test units, or expert opinions [17, Sec. 2]. It would also be important to establish the frequency or equivalent protocol of when to take measurements. Some literature also differentiates between reversible *Defects* and irreversible *Faults*, which [14, Sec. 4] recommends for them to be treated separately. Further useful information would include *Stressors* influencing the FMs.

A general characterisation of the *Maintenance Programme* development process would be to divide into either a deductive *Fault Tree Analysis* (FTA) based approach or an inductive *Failure Mode(s) and Effects Analysis* (FMEA) approach. An example process of implementing an effective FMEA is shown in [12, Fig. 1]. Survey results from industry participants, such as those in [14], often form the basis of industry-accepted *Condition Monitoring* techniques and frequencies. For example, the aforementioned [11] has an Annex dedicated to *Condition Assessment* metrics for specific *Condition Monitoring Techniques*, and a separate Annex dedicated to then developing a TAI for a TX. Leveraging the knowledge in these outputs can reduce development time and minimise unintended gaps in scope, however, they should be tailored to account for application-specific aspects such as costs or skill-availability.

Transformer Condition Monitoring Overview

There are too many *Condition Monitoring Techniques* to discuss each individually. Industry experience is highly valuable to determine which techniques are practically applicable prior to any costly investments. As per [22, Sec. 11]:
"It is certain that different users, different circumstances of use and different sizes and types of transformer will mean that there can be no one type of monitoring system to suit all transformers. Indeed the need for and type of monitoring required is likely to change during the lifetime of a transformer".

Furthermore, over time new information regarding specific TX designs are obtained through forensics and fleetwide trending that may guide future decisions. As per [15, Sec. 11.3], "experience has shown that several transformer design groups have inherent design weaknesses which reduce useful service life" and "transformer models periodically require updating (supported by evidence from forensic analysis) as further understanding of deterioration mechanisms is acquired during the transformer life cycle". [13, Fig. 2] tabulates a range of common *Condition Monitoring Techniques*

mapped against the FMs they can potentially detect. Table 2-3 is modified from Table 3 from [13, Sec. 8] to show only the FMs covered in Table 2-1.

Table 2-3: Transformer Faults matched to Measurement Parameters / Techniques

		Symptom or Parameter change or Detection Technique													
		Amps / Volts / Load													
		Visual													
		Oil Condition													
		Temperature													
		Partial Discharge													
		Dissolved Gas Analysis													
		Noise													
		Ultra-Sound													
		Vibration													
		PF / Tan Delta													
		Resistance													
		DFR / PDC / RVM													
		FRA													
		Excitation Current													
		Leak Reactance Flux													
Failure Modes															
	Failures Detected														
Dielectric Faults	Insulation Deterioration		●	●	●	●	●		○		●	●	●	●		●
	Moisture Ingress Content				●	●	●				●			●		
	Tap-changer Condition / Problem		●	●		●		○	●	○	●	○	●	●		●
	Oil Quality Deterioration				●		●				●			●		
	Arcing / Electrical Discharge						●		●	●	●	●		●	●	
	Connection Problem					●	●		●	○	○	●	●			
Thermal Faults	Overheating / Auxiliary Cooling System Problem								○		●		●	●	○	
	Low Oil Level								○	○	○		○	○	●	
	Oil Circulation System Problem										○		●	○	●	
Mechanical Faults	Winding Distortion	●		●	●				○							
	Winding Looseness							●	●	●						
	Core Looseness							●	●	●						
	Oil Leak														●	
	External Damage / Disturbance														●	
External Faults	e.g., Animals														●	
	Through Fault, e.g., lightning strike or short-circuit			●			●									
	Supply Faults, e.g., excessive harmonics / over fluxing															○
Performable Online?							✓	✓	✓	✓	✓	✓	✓	✓	✓	
Legend:	● = Likely Relationship	○ = Less Likely Relationship														

Source: Modified from Table 3 from [13, Sec. 8]

Appendix 8 in [14] has a more extensive set of tables where most FMs, *Components*, and applicable *Condition Monitoring Techniques* are mapped together alongside notes of alternatives and their relative sensitivity / interpretative power. Readers should be aware of evolving technologies and ensure its information presented from 2003 is still relevant. Table 1 from the more recent (2008) [22, Sec. 8] also provides guidance, recommending *Condition Monitoring Techniques*, and tabulating sensors applicable to

the main *Components* within the TX. Additionally, [22, Sec. 10] also lists alarm indicators for the system operator, as well TX *Alarms* and *Trip Contacts* which it states are "not normally thought of as part of a condition monitoring system... but still provide condition information". An even more recent (2015) example can be found in Appendix 2 of [8], which has a similar, extensive tabulation, mapping FMs, *Components*, and typical *Condition Monitoring Techniques*. However, it is focussed only on *Functional Failures* that can be characterised using *Continuous OLCM*. It focusses on the primary functions of the *Subsystems* and so is not as exhaustive in its scope.

Alternatively, the sensor choice for a given monitored parameter can be focussed upon. Table 6 from [21, Sec. 6] is an example tabulation of TX components against monitored parameters alongside common sensor types used for the application. As a side note, Table 1 of [21, Sec. 4] is also very useful; it links *Components* to their FMs and then their measurable symptoms, and lastly to the *Diagnostic Technique*. This rendition managed to summarise the topic sans the *Bushings* and *Tap-Changer* to a two-page table, whilst giving insight into most aspects. However, it lacks any detail regarding how the *Diagnostic Technique* can identify a given FM using the highlighted signals.

There is no consensus regarding what combination of *Condition Monitoring Techniques* to use. The Appendix of [9] notes their hydro powerplant engineers consider as a "sound basis for assessing transformer condition": DGA and FFA (for *Liquid Insulation*), Power Factor and excitation current tests, operation and maintenance history, and age. [21, Sec. 7] states "load, temperatures, dissolved gas-in-oil and moisture sensors" can constitute a "comprehensive monitoring system" that "can provide a major support to the operator when the transformer faces overload conditions". Note that the use-case for the assessments differ, leading to differing recommendations. These could be considered a lower baseline given other cited sources include many more techniques.

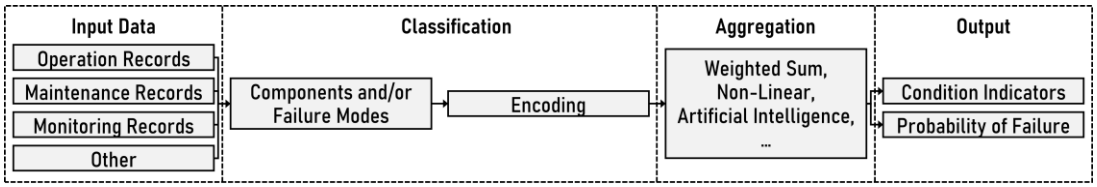
Transformer Condition Assessment Overview

Condition Assessment Scope

There are two primary challenges with *Condition Assessment* of TXs. The first is interpreting the *Condition Monitoring* data to arrive to a *Condition Assessment*, the second is consolidating the *Condition Assessment* data into a manageable volume that highlights the most needful TXs, typically as one or more TAIs. [7] is a Technical

Brochure on this topic with further details, especially Annex A and Annex B. Fig. 2-12 is a redrawn version of what is used in [7, Fig. C.1] to outline the problem scope. The first challenge is the process going from "Input Data" to "Classification", and the second is the process going from "Classification" to "Output" with an emphasis on "Aggregation". In addition, it can be more cost-effective for larger fleets to first apply a *Screening* to all TXs that may only indicate a potential problem, prior to investing into more sensitive tests for only the TXs that require it.

The terminology in the reviewed literature can sometimes be difficult to consolidate. For clarity, Fig. 2-13 diagrams the interpretation use in this thesis. This is based on the work from [7], [8], [14]. The lefthand column represents the typical life cycle of an asset. First an asset is *Commissioned* and installed, and then it begins *Operation* as described in Fig. 2-4 and Fig. 2-6. There may be *Routine Maintenance* carried out following a TBM protocol depending on the *Maintenance Strategy*. In addition, there may be *Condition Monitoring* ongoing to enable CBM and other more complicated protocols. *Condition Assessment* is chosen to encompass both *Screening* and *Detailed Interpretation*. One nuance neglected in Fig. 2-13 brought up by [14] is that there are three main contexts in which *Condition Assessment* is carried out: *Fingerprinting* during factory tests, evaluating or ranking assets, and *Fault Identification* at the site. These are sequential; factory testing occurs before or during the initial *Commissioning* as a proactive measure, evaluating assets happens during operation to better plan maintenance, and *Fault Identification* occurs after a *Fault* as a reactive measure. The outlined two-step process is most applicable to the second use-case of assessment; evaluating or ranking assets.



Source: Redrawn from [7, Fig. C.1], original from CIGRE © 2019

Fig. 2-12: Health Index Development Process

In this thesis, *Screening* encompasses the automated or routine *Condition Monitoring* information and its automated or routine interpretation for the purpose of identifying the assets that warrant closer attention. *Detailed Interpretation* is then that step. It is expected for the latter step to be manual, though it may be aided with automated tasks,

and that major investment decisions all pass through this validatory stage. The scope of the *Screening* or the automated tasks depends on the implementation. In general, it can include *Anomaly Detection*, *Fault Detection*, and *Fault Identification / Diagnosis*. The *Screening* will either indicate to take no action and to continue *Operation* as usual, or to inspect further via *Detailed Interpretation*. Specific known responses can be automated as part of a CMB protocol, such as ordering a consumable to be replaced during *Routine Maintenance*. During the *Detailed Interpretation*, no action is an option with the asset continuing *Operation* as usual, as is acquiring further information via modifying the *Condition Monitoring* protocol. This can be either altering the frequency of a test or introducing new tests, as shown on the righthand column of the figure. It may also be decided that an intervention is required, and depending on the evaluation, this may be in the form of either *Corrective Maintenance*, *Refurbishment*, or the *Scrapping / Decommissioning* of the asset. *Routine* and *Corrective Maintenance* aim to maintain the expected lifespan of the asset whereas *Refurbishment* aims to extend it, as shown on the righthand of Fig. 2-13.

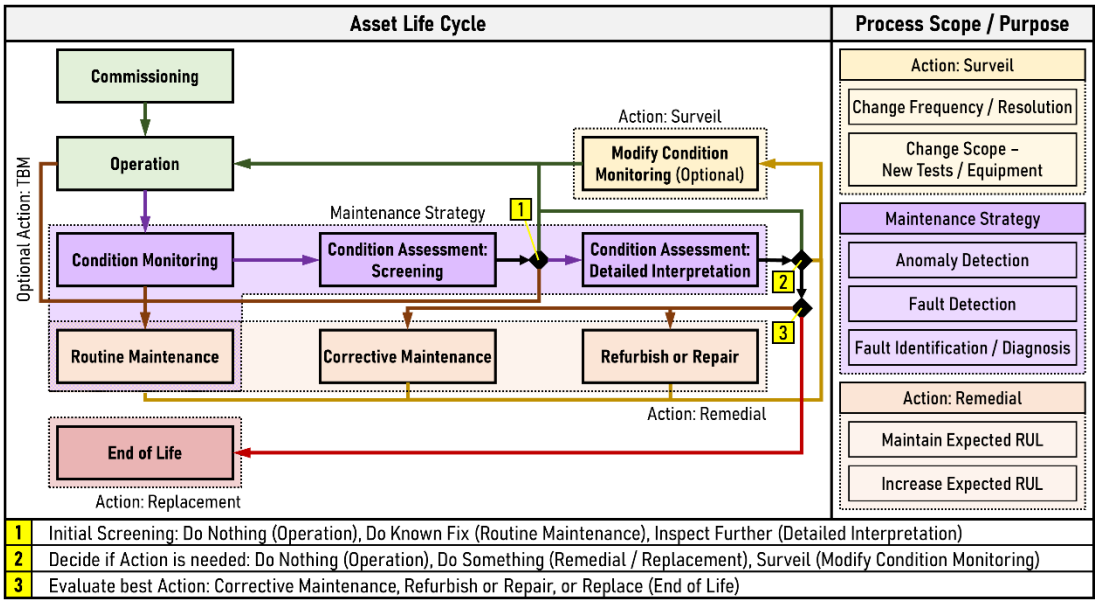


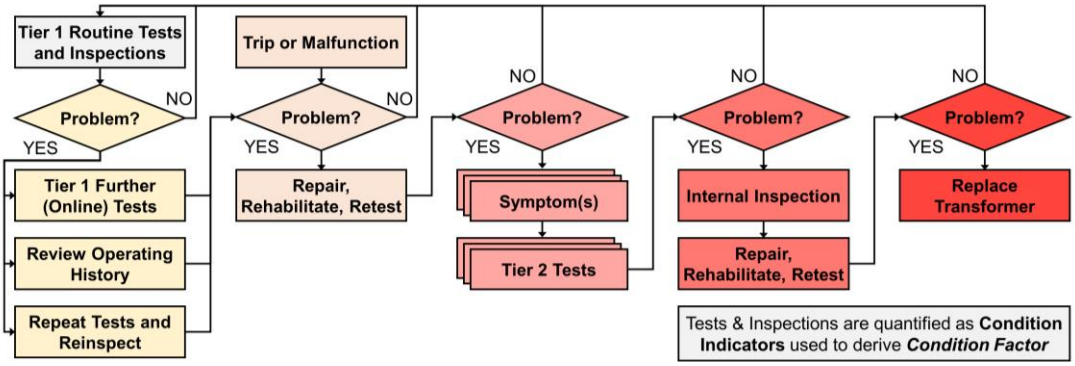
Fig. 2-13: Role of Condition Monitoring within Asset Life Cycle

Fault Severity is hard to define, as it can be part of both *Screening* and *Detailed Interpretation*, depending on context. *Fault Severity* within *Screening* is interpreted as a measure of how far the asset has deviated from the expected parameters whereas *Fault Severity* within *Detailed Interpretation* is interpreted as a measure of potential impact on either PoF or EoL. As an example distinction, if in one TX a gas is being produced at

a higher rate than in another, it may be flagged within *Screening* as having a higher *Fault Severity*. However, *Detailed Interpretation* may show that the composition of the gases produced in the other TX is indicative of a more severe *Fault Type* such as *Arcing* for example. In this case, the *Detailed Interpretation* would flag the other TX at a higher *Fault Severity*. In other words, considering how prevalent the symptom is compared to how important the symptom is when assessing severity.

Although the specific terminology differs across literature, those reviewed are broadly in line with the above interpretations or can be trivially mapped across. As an applied example, the Appendix of [9] outlines the programme used by USBR's hydro-plant engineers. A simplified version of [9, Fig. 25] and [9, Fig. A-1] is shown in Fig. 2-14, showing an initial *Routine Tests and Inspections* stage potentially triggering a selective range of additional tests ascending in scope. This is considered *Screening* as a first stage, followed by a multi-stage *Detailed Interpretation* process. An example of potentially confusing terminology would be in Fig. 2-6, modified from [8], which treats *Condition Assessment* as distinct to *Interpretation*. However, in full context, it is clear *Condition Assessment* aligns with *Screening*, recommending *Interpretation* only if prompted.

For TX DGA specifically, arguably the two most established standards are IEC 60599 [2] and IEEE Std C57.194 [1]. These are both detailed in Chapter 3. [2] aligns well with a two-step process as its *Diagnostics* is reserved for when there is an indication of a potential issue—much like an initial *Screening* prior to a *Detailed Interpretation* [2, Sec. 9]. Similarly, [1, Sec. 5] presents the terminology of *Detection*, *Evaluation*, and *Action*. As per [1, Sec. 5], "the interpretation of DGA data begins with the detection of an abnormal condition. When found, it should be followed by severity assessment and fault identification". This could be considered as the initial DGA samples being tested against limits is akin to *Screening*. As per [1, Sec. 3], *Screening* is defined as "a test protocol in which all transformers in a population are tested at regular time intervals (e.g., every year) to identify units which may require additional attention or remedial action. This protocol is used to identify transformers with potential fault activity".



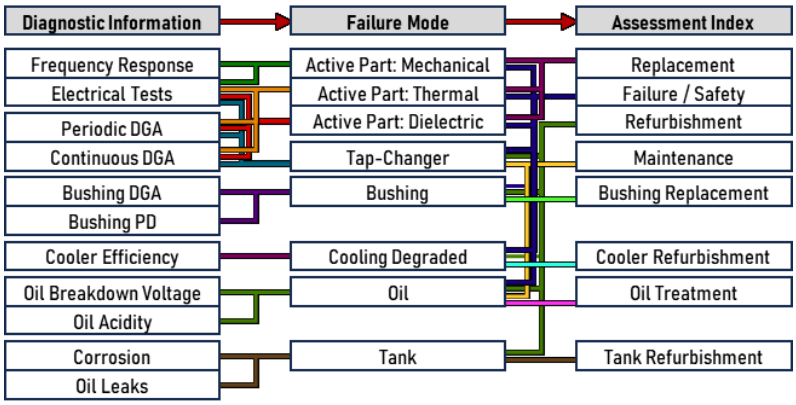
Source: Modified from [9, Fig. 25] and [9, Fig. A-1]

Fig. 2-14: Example Transformer Condition Assessment Methodology

Condition Assessment Indices Overview

The challenge of interpreting the *Condition Monitoring* data is too broad a scope to consider as it is almost wholly dependent on the *Condition Monitoring Technique*. However, in general, the derived output should be a reproducible, traceable value that is either quantitative or ordinal in nature. Key considerations for a TAI are outlined in [7, Sec. 2] and is paraphrased as stating that a scoring system should rank all TXs in a fleet such that those most in need of action or intervention are easily identifiable without masking any individual FMs requiring urgent attention. Additionally, that the scoring system should be reproducible and transparent / interpretable by any user, with reference to the purpose of the TAI.

Annex A of [7] provides a guidance on interpretation with examples. Consolidating these outputs consistently to produce sensible outputs without masking edge-cases needing attention can be challenging. Where resources allow, all flagged assets can be addressed in a timely fashion on an ad hoc basis. However, if there are pre-allocated budgets for investments that need spending, or a bottleneck in resources, then sequential ranking is needed. Even if capital is available, the pragmatics of personnel logistics must be considered, and so on. The ordering could be based on the potentially easier-to-quantify CoF. Assuming consolidation is required, using multiple TAIs relating to specific actions can be helpful and is often recommended. For example, [11, Sec. 3] states that a TAI should "preferably … be able to refer to the transformer needs in terms of replacement, refurbishment and maintenance". Fig. 2-15 redrawn from [7, Fig. 1-1] also indicatively demonstrates how *Condition Monitoring Techniques* can be linked to both FMs and specific TAIs. Although, [7, Sec. 1] warns it is an example only and that implementations should be derived from the available information.



Source: Redrawn from [7, Fig. 1-1], original from CIGRE © 2019

Fig. 2-15: Relationship of Diagnostics, Failure Modes, and Assessment Indices

Table 2-2 from [7, Sec. 2] outlines some methods to combine inputs with a brief description of their advantages and disadvantages. It is assumed that the *Condition Monitoring* data has been processed into a metric relating to a FM or specific *Condition Assessment*. Some methods are dependent on the metric. For example, if a reliable and accurate measure of PoF for each FM can be captured, then a combined PoF can be calculated. However, as stated in Annex 3 of [7], "there is no consensus in the literature regarding a methodology to assess the probability of failure". For example, Fig. 2-16 redrawn from Table B3 in [7] [7, Fig. B3] illustrates how *Condition Assessment* outputs for each category can be consolidated in different ways. The colour represents severity, and the "Simple Score" shows the count of the worst-case. A more expansive alternative would be to tabulate the count of each *Index* output. This could be summed but that can potentially mask issues. The example "Hybrid Score" combined the overall count with the worst-case category—although other hybrid approaches are available. Lastly, a non-linear score is shown using Equation (1):

$$TAI = \sum_{n=0}^{k-1} x_n \, i^n, \tag{1}$$

where i is equal to or greater than the number of FMs included in the TAI, x_n is the number of FMs per category, and k is the number of categories included in the FM assessment. As per Annex B of [7], "although it is not immediately obvious, the single numeric score per transformer indicates the timescale for action. A score above 81 can only be achieved if at least 1 sub-component needs urgent attention (i.e., is scored as Red)".

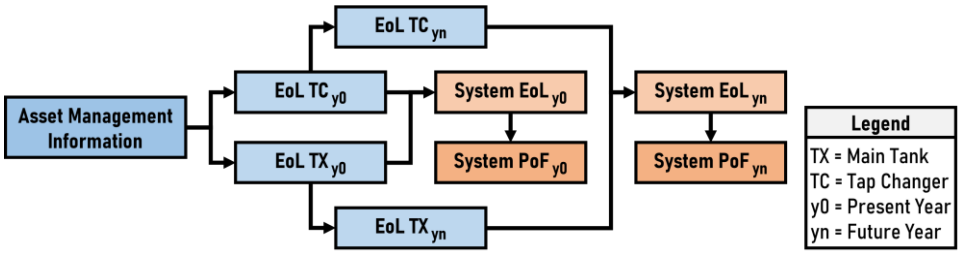
Transformer # :	1	2	3	4	5	6	7	8	9	10
Main Tank (Unrepairable)	P	G	O	O	R	Y	G	R	Y	O
Bushings	O	R	R	G	O	G	G	R	G	O
OLTC	O	G	Y	Y	O	G	G	O	G	Y
Red	0	1	1	0	1	0	0	2	0	0
Pink	1	0	0	0	0	0	0	0	0	0
Orange	2	0	1	1	2	0	0	1	0	2
Yellow	0	0	1	1	0	1	0	0	1	1
Green	0	2	0	1	0	2	3	0	2	0
Simple Score	1-P	1-R	1-R	1-O	1-R	1-Y	3-G	2-R	1-Y	2-O
Simple Summed Score	7	4	7	3	8	1	0	10	1	5
Hybrid Score	7-P	4-R	7-R	3-O	8-R	1-Y	0-G	10-R	1-Y	5-O
Non-Linear Score	45	83	93	13	99	5	3	171	5	21

Legend:		
Code	Colour	Value
G	Green	0
Y	Yellow	1
O	Orange	2
P	Pink	3
R	Red	4

Source: Redrawn from Table B3 in [7], original from CIGRE © 2019

Fig. 2-16: Example Derivations of Assessment Score

Annex B of [7] discusses *Weighting Factors* to account a component's relative importance. These can be made as complicated as needed to account for different FMs in different *Components*, although this can also obfuscate the raw data from the decision-maker. Conversely, they can be kept simple. For example, Fig. 2-17 redrawn from [15, Fig. 6] shows the UK regulator's "NOM" approach, which considers the TX *Tap-Changer* and *Tank* separately before recombining via equal weights.



Source: Redrawn from [15, Fig. 6]

Fig. 2-17: Probability of Failure Calculation for Transformers

One relevant aspect not directly addressed is the commonly stepped nature of *Condition Monitoring*, with widescale *Screening* prompting more detailed *Interpretative Assessment*. [7, Sec. 2] states that the more accurate indicator should take precedence where they overlap in scope. If the accuracies are similar, either averaging or taking the worst-case can be used, depending on the purpose of the TAI. Where indicators conflict, the worst-case result could still be used but should be flagged given the uncertainty around the assessment, requiring further investigation prior action. An example in [7, Sec. 2] is that DGA may indicate **PD** during *Screening*, which may then prompt the capture of the more accurate *Ultra-High Frequency* (UHF) measurements to confirm the presence and severity of the PD. The UHF would then supersede the DGA, with the latter simply increasing the confidence in the assessment it corroborates

rather than attempting to consolidate the outputs. This would mean that in the context of Fig. 2-16, there would assumedly be some data captured by *Interpretative Assessment* that would be missing during the initial *Screening*. In this case, the new tests representing measures of FMs would either confirm the poor status or override them with a lower level. However, keeping concise records can be complicated, leading to another key aspect of an effective TAI—ensuring adequate data quality. This includes conveying where data is missing or inaccurate, as well as confidence levels to provide the necessary context to the decision-makers.

2.3. Uncertainty Overview

2.3.1. Conformity Amidst Uncertainty

Taking a broad definition of *Conformity*, *Non-Conforming* results represent outcomes significantly inconsistent with predefined expectations. In *Conformity Assessment*, these expectations are *Specified Requirements* [29, Sec. 4.1]. It is generally assumed that if a *Non-Conformant* output cannot be repeated. it represents an anomaly not requiring action. However, if this occurs too frequently, it may indicate a broader issue with the *Condition Monitoring*. Generally, *Uncertainty* negatively correlates with the ability to interpret *Non-Conforming* results. The main challenge of *Asset Management* is arguably to cost-effectively account for *Uncertainty*. In its broadest sense, *Uncertainty*, as per [30, Sec. 2.2], is related to doubt. This thesis considers *Uncertainty* in CMA via three sources. First is the *Uncertainty* associated with an obtained value intended to represent a given metric. Second is the *Uncertainty* associated with limit(s) intended to signify relevant breakpoint(s) in interpreting said metric. Lastly, there is the *Uncertainty* associated with the metric's relevance to asset health. In practice, the commingling of factors is not so easily discretised.

Each source of *Uncertainty* can compound, and given inevitable constraints, priorities must be set to optimise reductions in *Uncertainty* based on functions of the expected time, cost, and impact. This echoes earlier discussions had in this Chapter, especially for *Condition Monitoring*. Similarly, as with FMEA, discretising the CMA process into stages helps identify potential sources of both *Uncertainty* as well as unexpected results. From the already surveyed literature, [11, Fig. 2.1]'s definition of *Transformer Intelligent Condition Monitoring* (TICM) is the closest to outlining this intended

process. [31, Ch. 10] also has a relevant definition in its "key steps in data acquisition and processing" figure [31, Fig. 10.1]. However, neither are ideal here as they neglect the period prior to CMA, i.e., design, manufacture, transport, and assembly of the TX.

Process	Potential Sources of Uncertainty & Unexpected Results	
	Knowledge-Based (Theoretical)	**Application-Based (Practical)**
Asset Realisation (Design & Manufacture & Installation)	▪ Incomplete understanding of: ▪ Design ▪ Materials ▪ Impact of Local Operating Conditions	▪ Natural variance due to: ▪ Manufacture ▪ Local Operating Conditions ▪ Issues due to: ▪ Defective Component(s) ▪ Transport & Installation
Data Acquisition (Monitoring Coverage & Sensor Data Capture)	▪ Poor Monitoring Choice: ▪ Poor Relevance to a Failure Mode ▪ Failure Mode (Mechanism) not Captured ▪ Poor Sensor Choice ▪ Inadequate Accuracy / Reliability / Sensitivity ▪ Unsuitable Working Conditions	▪ Issues due to: ▪ Defective Sensor ▪ Issue with Sensor Installation or Calibration ▪ Natural variance due to: ▪ Noise from Environment ▪ Sensor Data: Randomness / Noise / Missing
Data Analysis (Metrics Selection, Extraction, & Interpretation)	▪ Poor Metric Choice ▪ Poor Relevance to a Failure Mode ▪ Failure Mode (Mechanism) not Captured ▪ Poor Interpretation ▪ Incomplete understanding	▪ Poor Feature Extraction ▪ Error in Extraction Process ▪ Poor Interpretation ▪ Error in Interpretation Process

Fig. 2-18: Potential Sources of Uncertainty and Unexpected Results

Fig. 2-18 frames the discussion using a different process structure to tabulate some potential sources of *Uncertainty* and unexpected results. It highlights the complexity of determining the source of an unexpected result during CMA. Fig. 2-18 divides the sources broadly into two categories: *Knowledge-Based (Theoretical)* and *Application-Based (Practical)*. The *Theoretical* class represents *Uncertainties* caused by design. For example, a decision to neglect a particular FM is an intentional source of *Uncertainty*. However, even if said FM is neglected as an oversight in programme design, it remains a (lack of) *Knowledge-Based (Theoretical)* issue. Another example are the metrics chosen to represent *Condition Monitoring* data or to quantify the levels of degradation. In contrast, the *Practical* class relates to consequences of actions and implementations. For example, errors in quantifying even an ideal metric may introduce *Uncertainties*.

Each of these, though most likely not all concurrently, can contribute to *Uncertainty* and unexpected results. It is therefore helpful to ascertain the most likely source of an unexpected result prior to committing resources to staging an intervention. This could be approached conceptually like a FMEA: if the common 'symptoms' of an error caused by a specific stage in the data processing methodology can be determined, then it may be possible to attribute detected errors to said stages. This is of course challenging practically. Nevertheless, such models can also be informative regarding *Uncertainty*

associated with a value. Focussing on this second use-case, Appendix E of [32], characterises this approach as *Cause-and-Effect Analysis*. It describes the general Ishikawa (fishbone) diagram as a "hierarchical diagram that shows how multiple effects accumulate to influence an outcome" [32, p. 80]. The taxonomy can differ without invalidating the approach, for example, Appendix E of [32] discusses the *5M Method* and the *Measurement System Analysis* (MSA). These are generally interchangeable and serve as conceptual tools to either better illicit knowledge from the implementer and organise the information for others.

Asset Realisation

As discussed with Fig. 2-4, there are many potential sources of unexpected results prior to the *Operation* stage, these are termed as the *Inherited Condition*. These can similarly be sources of *Uncertainty*. For example, manufacturing tolerances can lead to a range of expected values. This stage of the process is here termed as *Asset Realisation*. There will be *Uncertainty* related to the theoretical understanding of the design. For example, older TXs were designed without modern software modelling and some complex properties such as the specific location of the hotspots may be uncertain. Similarly, the degradation features of every component and material will not be known. For example, some passivators were found to harm components within a TX over time [7, Sec. 12]. There is then another layer of *Uncertainty* introduced in the manifestation of the design, i.e., an asset's construction. There will be manufacturing tolerances within materials and an expected range deemed acceptable for many parameters, however the specific permutation of the possibilities for each asset is generally not known. This again is a bigger issue with older assets as manufacturing knowledge and context of the specific designs can get lost over time.

Data Acquisition

The stage termed *Data Acquisition* also includes the *Uncertainty* from the decisions made regarding the *Condition Monitoring* coverage. A physical asset is not typically inspected in its entirety at every instance. Instead, from a function of cost, availability, and understanding of relevant FMs, a selection of *Condition Monitoring Techniques* is developed into a *Condition Monitoring Programme*. The *Condition Monitoring Programme* must specify how the representation of the asset is captured. In some cases, these representations may be poorly correlated with *Faults* or not conclusive.

Alternatively, it is also possible for a FM to be missed entirely or simply ignored due to being deemed irrelevant. The choice of sensors, their intended installation locations, and their actual installation, are all also potential sources of unexpected results. An unsuitable sensor selection for a given application may result in, for example, insufficient sensitivity or excessive noise due to a lack of adequate EM shielding. For example, [33, Sec. 9] hypothesised that a particular DGA online monitor had particularly poor H_2 accuracy due to is use of a "solid state sensor not very adapted for that purpose". An appropriate sensor may still be installed incorrectly or have incomplete coverage, or the sensor itself might have an issue due to damage or poor calibration. Lastly, even a functioning sensor may output anomalies due to random noise. This aspect of the *Data Acquisition* stage aligns with either the "Input from sensors, IEDs, Transducers" or "Data Acquisition" from [11] and [31], respectively, for further reference if desired.

Another important aspect is whether the data acquisition process is automated or manual. This is a facet neglected in the other two ([11], [31]) pieces of literature that instead focussed on automated processes. However, this is particularly relevant for DGA where traditionally, the gas samples were extracted manually and then sent to a laboratory for manual analysis. Although, more recent approaches sometimes have an automated extraction and analysis process. These manual interventions are also potential sources of *Uncertainty* when applicable. To minimise the *Uncertainties* here, the *Operator Sampling Procedure* should be well designed with little ambiguity. Then, the *Operator* should be adequately trained according to said procedure. Even then, *Operator* adherence to the procedure remains a factor. [34, Sec. 4] characterises *Uncertainties* attributed to this source as either *Operator Error* or *Operator Skill*, citing the use of a stopwatch to measure time as an example of the latter: there is a variable amount of *Uncertainty* associated with the measurement related to the *Operator*'s reaction time that is 'normal'.

However, it can be challenging to differentiate issues due to *Operator Error*, *Operator Skill*, or a poorly designed *Operating Sampling Procedure*. One heuristic is that if repeating the measurement process eliminates it, then perhaps it is more likely to be due to *Operator Error* which is assumed to be an infrequent occurrence [34, Sec. 4]. However, if it appears *Operator Errors* are high, this indicates a potential issue with the

Operating Sampling Procedure. For example, it may be unrealistic in its demands of the *Operator* given the working conditions or contain too many subjective steps. (It may also indicate an issue with an individual *Operator*). Similarly, an *Operating Sampling Procedure* that unnecessarily relies too much on *Operator Skill* is poorly designed. For instance, relying on an *Operator* to measure and record from an analogue dial far from eye-level increases the chance of parallax errors.

Any applicable analogue-to-digital signal processing and initial data transfer to the processing destination are included in this stage in Fig. 2-18. This would include cases where data is transferred manually. For example, if an *Operator* reads the instrument and then writes it down; there is the potential for typographical errors. For DGA specifically where the oil sample is often extracted then sent to a laboratory, there is potential for the sample to be affected in the process.

Data Analysis

The sensors (or *Operators*) capture data that must ultimately be converted into a decision, even if the decision is to take no action. This requires some form of metric to which a limit or similar can be compared against. This stage termed *Data Analysis* is focussed on extracting the relevant metrics from the data and interpreting them. Often the data is compressed or transformed in some way. It is possible that the summarising metric loses valuable information related to the FM. Also, the metrics or features are often extracted / calculated automatically via methods such as signal processing. It is possible for this process to fail, especially if there is noise present. Similarly, there is some uncertainty regarding interpretations. Another more fundamental aspect is the knowledge behind the decision making. For example, not all *FM Mechanisms* are known or documented correctly. Data validation or qualification is assumed to occur at this stage as the preliminary *Data Analysis* step. However, this is not always clear-cut; for example, an *Operator* extracting DGA samples may check for contamination.

2.3.2. Causes of Uncertainty in Data

It is stated in [7, Sec. 4] that all assessments include unavoidable levels of *Uncertainty* due to imperfect assessments and potentially unpredictable degradation progression. For cases where the data is available, [7, Sec. 4] highlights three sources of *Uncertainty*:

- Incorrect data entry, or erroneous, or questionable test results
- Uncertainty in the condition assessment
- Aged data.

Incorrect data entry, or erroneous, or questionable test results

Uncertainties from these sources can arise at any of the stages in Fig. 2-18. For example, design specifications may be incorrectly recorded during the *Asset Realisation*, or the wrong interpretation selected during *Data Analysis*. Errors are most likely introduced during *Data Acquisition* and identified during *Data Qualification* within *Data Analysis*. Some use-cases for different *Data Validation Techniques* such as double entry, range checks, etc., are outlined in [7, Fig. 4.2]. These can be effectively combined to broaden their scope and are among the most cost-effective measures for reducing *Uncertainty*.

Within the context of *Measurement Uncertainty*, these may be attributed to what [34, Sec. 4] terms as *Operator Error*: considered by [34, Sec. 4] as out of scope for *Measurement Uncertainty* and instead it recommends simply repeating measurements as an effective approach to mitigate its impact. If symptomatic, they may be flagged as statistical *Outliers* or *Anomalies*. Within the methodology for evaluating *Measurement Uncertainty* outlined by the ISO 5725 series [35], [36], [37], [38], [39], [40], an *Outlier* is defined as a "member of a set of values which is inconsistent with the other members of that set" and would be removed prior to further analysis. These are termed *Blunders* in [30, Sec. 3] where it states "large blunders can usually be identified by a proper review of the data; small ones could be masked by, or even appear as, random variations. Measures of uncertainty are not intended to account for such mistakes".

Uncertainty in Condition Assessment

It is stated in [7, Sec. 4] that *Condition Assessment Techniques* vary in 'accuracy'. More 'accurate' but costly methods are sometimes reserved for cases where cheaper methods first indicate a problem. A relevant example in [7, Sec. 2] is how though **PD** is detectable by DGA in a TX, to locate the **PD** and determine its type, more 'accurate' sensors such as *Ultra-High Frequency* (UHF) sensors are needed. The nuance here is that when

compared to the context of *Measurement Uncertainty*, the latter would generally assess how *Accurate* the measurement of gas concentrations in DGA were, or how *Accurate* the measurement of frequencies in UHF were: it does not inherently extend to either how accurately they detect **PD** or locate its source.

It is also highlighted in [7, Sec. 4] that manual inspections are subjective, and clear protocols can help minimise *Uncertainty*. Calibration training, along with the use of qualified inspectors and third-party providers, are potential mitigative measures. Autopsies and lab-based testing can also help improve the fundamental understanding of an asset which can help improve the certainty in which data is interpreted. *Uncertainties* from these sources would fall under the *Data Analysis* stage in Fig. 2-18.

Aged Data

A data point represents a snapshot into the condition of the asset at that point in time. As time progresses, so could the degradation within the asset. As the duration of time increases, so does the opportunity for the condition of the asset to deviate away from what it was when captured. However, [7, Sec. 4] notes that this represents only the **potential** for data to lose relevancy—old data may still reflect the asset if conditions remain unchanged. Unfortunately, there is no way to know for sure without getting a new data point to check. Lower value assets often have less data available for *Condition Assessment*, leading to a higher level of *Uncertainty*. An example is provided in [7, Sec. 4] of how two DGA samples of a TX showing low absolute values of *Key Gases* with no change could allow a reasonable assumption that the latest test result is unlikely to become obsolete within the next year. However, if the two DGA samples shows a large change, the *Timeliness* period would shorten significantly, and the chance of *Obsolescence* one year later is high. According to [7, Sec. 4], *Timeliness* refers to the expected timespan for which the data is assumed relevant, and *Obsolescence* refers to the extent to which the data is representative of the current state.

Another challenge with older data is the potential loss of context over time. For example, the network may have changed. Similarly, sensor lifespans are often shorter than some assets such as TXs. Changes in sensors, or technology-related compatibility issues over time become more likely, especially where proprietary solutions are being

relied upon [16, Sec. 10]. If these aspects are not all captured and presented alongside the data, the wrong assumptions may be made.

The concepts of *Timeliness* and *Obsolescence* can be difficult to incorporate in typical *Measurement Uncertainty*. Generally, *Measurement Uncertainty* is a static evaluation: it may be updated considering new information, but it does not inherently change. This is not to say these two concepts cannot be incorporated into *Measurement Uncertainty*. For example, via *Bayesian Belief* methods, expected to be covered in the currently unpublished Part 8 of GUM. The context in which time **is** incorporated into *Measurement Uncertainty* is regarding the expected *Accuracy* of results. It is generally modelled such that measurements taken in quick succession, ceteris paribus, are expected to have higher *Precision* than those taken over longer intervals. Annex C of Part 6 of GUM [32] discusses this aspect, including reference to the ISO 5725 series for what it calls a "top-down" approach. Another approach is if for example, a measurand is known to decay over time, then the time elapsed until measurement may then constitute as a bias, but more typically the impact on the bias is not known. Part 6 of ISO 5725 [39, Sec. 6] has a Section on the related topic of *Stability*, although this is not explored further in this thesis.

Missing Data

A distinction is made by [7, Sec. 4] between cases where data is available and where it is not, i.e. missing data. It could be argued maximum *Uncertainty* is when a data point is missing entirely. Generally, either the data was never captured, or it was lost or irreversibly corrupted in the process. Referencing Fig. 2-18, some examples include:

- Sample is simply pending until its due collection date.
- *Operator* cannot get access to site or otherwise make the measurement.
- Sample gets contaminated, or data gets corrupted, for example, an issue with the sample container or sensor.
- Sample / data lost in transit, for example, an issue with mail or communication links.

It is stated in [7, Sec. 4] that, where reasonable, a TAI should still function despite missing data, however, a minimum amount of data should be required to ensure a reasonable output. The nature of the missing data is a critical aspect to ascertain to correctly address it. Depending on context, missing data may be best left "missing" or

imputed. Alongside [7, Sec. 4], [41, Sec. 1], [42, Sec. 1] can be referenced for further details on characterisations of missing data, as it is not explored further in this thesis.

2.3.3. Incorporating Uncertainty Overview

It is stated in [7, Sec. 4] that *Uncertainty* should be conveyed with a TAI to better inform the decision-makers on the relevant context. However, it is important to understand the concept of *Uncertainty* can differ between applications, as can the method used to estimate it. This Sub-Section provides only a general overview to convey the point. Two broad categories are suggested in [7, Sec. 4], one is creating a separate *Index* to convey *Uncertainty* information, and the other is attempting to integrate the *Uncertainty* into the TAI's output. The methods need not be complicated nor difficult to implement. For example, [7, Sec. 4] suggests including an indicator wherever data was missing as a simple method to inform the decision-maker. If creating a separate *Index*, [7, Sec. 4] includes an example where a *Data Quality Index* is created based on how recently the data was collected and its perceived reliability. Another example was a *Completeness Index* that was the percentage of missing data compared to the total expected data.

If attempting to integrate *Uncertainty* into the TAI, a simple method is to output a range of values such as the minimum, maximum, and expected value. However, this can sometimes be difficult to interpret when *Uncertainties* are high and almost the whole range of outputs are seemingly possible. This motivates a weighting mechanism to emphasise more likely scenarios, or similarly, a mechanism to curtail the range by excluding the least likely scenarios. The natural method is arguably the use of a probability distribution or equivalent for discontinuous data. This can be combined with the idea of outputting a *Coverage Interval* to provide an output range expected to cover the "true value" to a given *Coverage Probability* corresponding to an expected given percentage of cases. This topic is discussed more thoroughly in Annex G of [30]. A challenge to implementing a probabilistic approach is that the output probability distribution needs to be specified. As this distribution represents outcomes that did not necessarily happen, it can be challenging to calculate. A potential alternative is to instead specify the probability distribution(s) of the input and propagate them through an *Uncertainty* model on the assumption that the individual inputs are easier to specify. The outputs of the *Uncertainty* can then be calculated either analytically or estimated via numerical computation. Although analytical methods provide exact results and can

explicitly convey the influence of variables, they can be impractical in cases with too many variables, complex processes, or atypical probability distributions [43, Sec. 7].

Two general categories of estimation methods being highlighted here are analytical simplifications, and sampling approaches. The general motivation for analytical simplification is to allow for the calculation of otherwise overly complicated cases within a reasonable degree of accuracy. A typical example is to assume a *Gaussian* ($\mathcal{N}$) distribution as representative of the unknown empirical distribution. This is explained further in Annex G of [30] and underpinning it is the *Central Limit Theorem* and its implied consequence that combined distributions will, very generally, converge towards a $\mathcal{N}$ distribution. Another common example relevant here are approximation functions for integrating, such as the *Riemann Sum* [44]; these are especially useful where an empirical distribution may be known but not its analytical function.

Sampling approaches instead typically infer estimates based on outputs of repeated trials or simulations. Different sampling strategies may be employed, most typically either random or deterministic. The former can also be called the family of *Monte Carlo Methods* (MCM) [45, Sec. 4]. The latter is sometimes referred to as the family of *Quasi-Monte Carlo Methods* [46]. In general, deterministic sampling attempts to concentrate samples near points of interest, for example, where the output probability distribution appears to change the most. The published precursor of the work presented in Section 5.2, [47], uses another example, where samples are taken at pre-determined intervals along the *Cumulative Distribution Function* (CDF). A CDF is simply a cumulative sum of a probability function and thus provides the aggregate probability of obtaining values less than, or equal to, a given limit. This concept, along with a suggested procedure, is explained in Appendix D of [45]. The advantages of MC-like strategies are that they are broadly applicable even when the analytical solution is intractable or too complicated and time-consuming to solve [7, Sec. 4], [48, Sec. 7]. They also do not require as many specific assumptions to be met as analytical simplifications. However, they can also be computationally intensive and themselves may need simplifications to allow for reasonable runtimes. Furthermore, these two general categories of estimation methods are not necessarily mutually exclusive, and often analytical simplifications are made within an MCM.

This thesis explores propagation of *Uncertainty* probabilistically, but it is recognised that it is not the only valid approach. In Sub-Section 2.2.3, some of the *Interpretation Techniques* outlined by [11, Sec. 4] were discussed. These would again be relevant here as said techniques can sometimes either be adapted to accommodate *Uncertainty* or do so inherently. For example, [49] used a *Data-Driven* AI-oriented approach using a *Random Forest* model to estimate a TAI which incorporated a bespoke measure of 'certainty' to weight components designed to handle *Uncertainty* due to missing data. Variants such as *Quantile Regression Forests* have also been shown capable of estimating *Uncertainty* in its probabilistic context [50]. An example of a *Data-Driven* approach that can inherently accommodate *Uncertainty* is a probabilistic *Gaussian Bayesian Network* (GBN) or similar *Markov Tree*. A GBN is included in [51] in its ensemble intended to perform DGA-based TX *Diagnosis* where the focus was on *Interpretation Uncertainty* regarding the thresholds. The conceptually similar *Markov Tree* is used in [52] but structured slightly differently, where the focus was instead on *Measurement Uncertainty*. It is argued in [51, Sec. III] that GBNs can "capture causality among random variables (RVs) and infer uncertainty information", citing [53]. An overview of *Bayesian Statistical Models* is also provided in [32, Sec. 11], stating that they "reflect an understanding of uncertainties associated with both inputs and outputs as characterizing states of incomplete knowledge about the true value of the input quantities and of the measurand".

An example *Knowledge-Based* approach is found in [54], which uses a *Rule-Based Expert System* for DGA-based TX *Diagnosis*, incorporating *Interpretation Uncertainty* via *Fuzzy Sets* applied to thresholds suggested by established methodologies. The concept of *Fuzzy Sets* is an application of *Fuzzy Logic* popularised in the seminal paper, [55]. The general principle being having gradated and potentially overlapping thresholds rather than binary ones. At the risk of oversimplifying [56, Sec. 1]'s explanation, if probability theory considers the likelihood of a belief being 'correct', *Fuzzy Logic* instead considers the extent the belief is 'correct'. Clearly, there are conceptual similarities, and they can be complementary approaches as argued by the original author of [55] in [57]. *Bayesian Belief Networks* and *Fuzzy Logic* are empirically compared in [58] using an example in human reliability analysis, where it concludes that *Fuzzy Logic* can be more transparent in its inputs and outputs in cases with very limited available knowledge, but that its interpretation can be less intuitive, potentially

requiring an intermediate process called *Defuzzification*. A more comprehensive overview of *Fuzzy Logic* applied to TX *Fault Diagnosis* is provided in [59].

2.4. Transformer Dissolved Gas Analysis

2.4.1. General Principles

As per [1, Sec. 4], "dissolved gas analysis (DGA) is the identification, measurement, and interpretation of the gases dissolved in the insulating liquid". DGA is also referred to as *Dissolved Gas-in-Oil Analysis*, and it is a well-established *Condition Monitoring Technique* with the potential to identify a wide range of FMs relating to a TX's *Active Part*. [7, Sec. 1], states it is the "industry standard for the detection and determination of faults in TX" and that it is "recognized worldwide as the main tool to prevent failures of power TXs". Similarly, [9, Sec. 6] states that DGA is "by far, the most important tool for determining the health of a transformer", with [21, Sec. 5] explaining it is often compared to a "blood test in its diagnostic value".

This thesis assumes mineral oil as the *Liquid Insulation*. Captured within the oil, gases generated by various processes within the TX can be quantified and analysed to infer the state of the TX. *Solid Insulation* and oil both generate gases when degrading. The specific gases and their quantities will depend on the mode of degradation, principally affected by the ensuing temperature and energy involved. Carbon monoxide (CO), carbon dioxide (CO_2), oxygen (O_2), and water (H_2O) are released from the cellulosic materials used in *Solid Insulation*—paper and pressboard [9, Sec. 6]. The oil can release hydrogen (H_2), methane (CH_4), ethane (C_2H_6), ethylene (C_2H_4), and acetylene (C_2H_2) by its degradation [9, Sec. 6]. Lastly, but still important to consider, is the atmosphere as a source of gases, from which CO_2, O_2, N_2, and H_2O can all be absorbed [9, Sec. 6]. Other gases may be present but as per [1, Sec. 4], they are "ordinarily ignored for transformer DGA" such as argon, propane, and propylene.

Intuition can be gained through two related perspectives, the first being the standard enthalpies of formation for the gases and the second being the indicative, relative quantities of generation at varying temperatures. Regarding the first, [1, Annex F], [3], [4] explain how certain gases, such as C_2H_2, require more energy to form via the decomposition of mineral oil, and thus can be indicative of a higher energy event such as *Arcing* (**D**). The enthalpies of formation for the gases are shown in Table 2-4, where

the sources for the gases are via N-Octane for a mineral oil proxy, and the cyclic form of glucose for a cellulose proxy. As per [4, Sec. C], "the enthalpy of formation of substance B from substance A is the amount of energy required to produce one mole of B from A...". The values of all gases, except CO_2, were obtained from Table 6 in [3, Sec. D], CO_2 was sourced from [1, Annex F]. However, note that this is a simplification of the overall process. For example, [2, Sec. 4] states that in addition to high temperatures, C_2H_2 also requires a "rapid quenching to lower temperatures, in order to accumulate as a stable recombination product". Therefore, this perspective alone does not capture all relevant aspects to DGA.

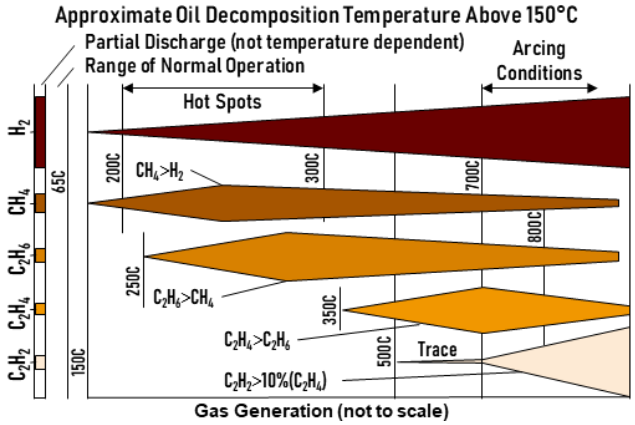
Table 2-4: Enthalpies of Formation of Fault Gases

Gas Name	Chemical Formula	Enthalpy (kJ/mol)
Methane[a,1]	CH_4	77.7
Ethane[a,1]	C_2H_6	93.5
Ethylene[a,1]	C_2H_4	104.1
Hydrogen[a,1]	H_2	128.5
Acetylene[a,1]	C_2H_2	278.3
Carbon Monoxide[b,1]	CO	101.4
Carbon Dioxide[b,2]	CO_2	30.2

Note. a: formation from N-Octane
b: formation from Cyclical Sugar Molecule (Cellulose Proxy)
Source: Values from 1: Table 6 of [3, Sec. D], 2: [1, Annex F]

The second useful perspective is the approximate gas generation via oil decomposition at varying temperatures, shown in Fig. 2-19 redrawn from [9, Fig. 48] which attributes [60], [61]. This can help explain why the ratios of gases can be insightful to differentiate *Fault Types*. A summarisation of the *Fault Types* described in Annex C of [1] is provided in Fig. 2-20. Moving away from the energy source, the volume of oil affected increases, and the temperature decreases as the heat is dissipated. Therefore, a range of gases may be produced by a single event. Additionally, H_2, CH_4, and CO are also produced by normal ageing [9, Sec. 6]. It is important to stress Fig. 2-19 is indicative rather than definitive. The similarly scoped Fig. 2-21 from [1, Fig. 1] is stated to only "schematically illustrate" the concept and shows slight nuances in the H_2 generation for example. Note that in Fig. 2-21, A*rcing of Low Energy* (**D1**) is placed as having a higher temperature than that of A*rcing of High Energy* (**D2**). This is because though the net energy released is lower in **D1**, it is released over a very short duration, creating a concentrated hot spot. In contrast, **D2** lasts a longer time, allowing its greater energy release to be dissipated. This can 'dilute' the gases generated at its hottest point with those generated

around it and means that its average temperature is lower [62, Sec. 2]. Regarding *Partial Discharge* (**PD**), as per [2, Sec. 5], there are generally two causes, the *sparking-type* and *corona-type*. The first are small arcs, similar to **D1**, occurring in the oil or paper phase whereas the second occurs in a gas phase, for example, in gas bubbles or voids. These are what is referred to as **PD** in Fig. 2-21.



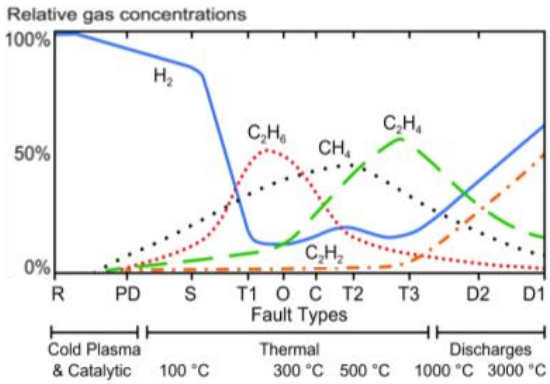
Source: Modified from: [9, Fig. 48] which attributes [60], [61]

Fig. 2-19: Relative Gas Generation in Mineral Oils

- **PD**: Corona Partial Discharges
- **O**: Overheating (T < 250C)
- **S**: Stray Gassing of Oil (T < 200C)
- **T1**: Thermal Faults (T < 300C)
- **T2**: Thermal Faults (300C < T < 700C)
- **T3**: Thermal Faults (700C < T)
- **DT**: Mixed Electrical and Thermal Faults
- **D1**: Low Energy Electrical Discharges
- **D2**: High Energy Electrical Discharges
- C: Hots Spots w. Paper Carbonisation (300C < T) – 80% Probability of Carbonisation
- Not Determined

Source: Summarisation based on Annex C of [1, Annex D.4]

Fig. 2-20: Summarised Fault Type Definitions



Note: Legend provided in Fig. 2-20

Source: Modified from: [1, Fig. 1]

Fig. 2-21: Relative Gas Generation in Mineral Oils

The topic of interpretation is open-ended, as per [9, Sec. 6]: "there are few, if any, 'cut and dry' DGA interpretations; even experts disagree". There are general *Fault* causes associated with the gas generations, but there are also other factors such as the TX's

load and ambient temperature. One approach termed the *Key Gas* method, as shown in Table 2-5 retabulated from Table D.1 from [1, Annex D], focusses on specific gases associated with a *Fault Type*. Although [1, Annex D] warns that even "when applied manually by experienced DGA users, the number of wrong *Fault Identifications* with Key Gas method is lower (typically 30%) but still high". Table 2-6 retabulated from Appendix C.4 of [62] is more recent example of a similar concept and includes the relative severity of the *Fault Types*. The definitions of the *Faults* are listed in Fig. 2-20. Appendix C.3 of [62] states that **D2** *Faults* and those occurring in paper (**C**, **D1** in paper) should be prioritised.

Table 2-5: Key Gas Method

Key Gas	Fault Type	Typical Proportions of Generated Combustible Gases
Ethylene (C_2H_4)	Thermal mineral oil	Predominantly Ethylene with smaller proportions of Ethane, Methane, and Hydrogen. Traces of Acetylene at very high fault temperatures.
Carbon-Monoxide (CO)	Thermal mineral oil and cellulose	Predominantly Carbon Monoxide with much smaller quantities of Hydrocarbon Gases. Predominantly Ethylene with smaller proportions of Ethane, Methane, and Hydrogen.
Hydrogen (H_2)	Electrical low energy partial discharge (PD)	Predominantly Hydrogen with small quantities of Methane and traces of Ethylene and Ethane.
Hydrogen & Acetylene (H_2, C_2H_2)	Electrical high energy (arcing)	Predominantly Hydrogen and Acetylene with minor traces of Methane, Ethylene, and Ethane. Also, Carbon Monoxide if cellulose is involved.

Source: Retabulated from Table D.1 from [1, Annex D]

Table 2-6: Severity of Types of Faults or Stresses

Fault Type	In Paper		In Oil	
	Main products formed	Severity	Main products formed	Severity
D2	C, C_2H_2	Very High	C_2H_2, C	Very High
D1	C, C_2H_2	Very High	C_2H_2, C	Moderate
T3	C, C_2H_4	Very High	C_2H_4, C	Moderate
T2	C, CH_4	High	CH_4	Low
T1, O	C_2H_6, CO	Moderate	C_2H_6	Very Low
Corona PD	H_2	Low	H_2	Very Low
S, T < 700 ºC, Ageing	CO_2, Furans, Alcohols, Low DPs of paper	Very Low	H_2	Very Low

Source: Retabulated from App Table C.4 from [62], original from CIGRE © 2019

2.4.2. Gas Level Interpretation

Although the general concepts are well-established, there is no agreed upon method to determine whether a *Fault* is present. In general, the absolute or changes in gas levels can be considered, and for both, either each gas individually or combined. It may also be possible to use a combination of both absolute and changes in levels.

Absolute Gas Levels

An advantage of using absolute gas values is that it requires only a single sample to analyse. Earlier implementations often used *Total Combustible Gas* (TCG) which was thought to be one of the most important indicators [9, Sec. 6]. TCG is a simple sum, allowing for reference to a single value rather than each of the gases: H_2, CH_4, C_2H_6, C_2H_4, C_2H_2, and CO. This metric was previously included in Standards such as the earlier versions of [1]. Although, it is now relegated to the "Historical Material" Annex in the newest version. *Total Dissolved Combustible Gas* (TDCG) is a related term that is the gas dissolved in the oil as opposed to in the headspace [63, p. 33].

Two criticisms of TCG are provided by [4, Sec. B], the latter being referenced from their earlier paper, [3]. The first is that the included H_2 and CO are not exclusively *Fault* related and make up most of the metric. The second is that CH_4 and C_2H_6 are treated equal to C_2H_2 and C_2H_4 whereas the first pair are associated with low and medium-range thermal *Faults* and the second pair are associated with the more serious arcing and high-range thermal *Faults*. These can be condensed into a single argument; that it is inappropriate to summate these different gases whilst neglecting their different implications. An alternative weighted summation is proposed in [3], [4], where the weights are inspired by the values shown in Table 2-4, here termed the NEI method. Another example approach is here termed the *Lapworth Scoring Algorithm* (LSA) method published in [5] that has a derived version used in industry. It also criticises TCG, and proposed weightings based on the relative significance of a gas. In addition, it attempts to scale the values using CH_4 as a denominator to mitigate the variations in designs and ages in TXs. NEI and LSA are discussed in more detail in Chapter 3.

The current publications of both [1], [2] instead focus on each gas individually for assessment. These are explored further in Chapter 3. Applying percentile-based limits to each gas individually will, in a naïve probabilistic sense, increase the likelihood of a sample being flagged which can itself be an issue. This is brought up in [5, p. 139] and readily acknowledged in [1, Sec. 5]. Aspects like loading should also be considered. For example, [64, Sec. 5.2] highlights that typical TCG values in UK are lower because it is standard procedure to operate at 60% of nominal load. Rather than a generic limit for absolute gas levels, [62] instead suggests limits for specific *Fault Types*.

Relative Gas Levels (Ratios)

For *Fault Identification*, many methods rely on ratios of gases rather than absolute values. As per [21, Sec. 4], "gas concentration ratios are thus a more reliable indication of an incipient problem than individual gas concentrations". An earlier example is the *Doernenburg Ratio* method, although as per [1, Annex D.2], it is "a historic method less used today". *Rogers Ratio* method was an evolution the method, simplifying from five key gases to three key gases [9, Sec. 6]. It is stated in [9, Sec. 6] that Fig. 2-19 was "used by R.R. Rogers of the Central Electric Generating Board (CEGB) of England to develop..." the method. It relies on three sets of ratios: C_2H_2/C_2H_4, CH_4/H_2, and C_2H_4/C_2H_6. Arguably, the ratios provided in the Table 1 of [2, Sec. 5] would similarly supersede *Rogers Ratio* method as it had superseded *Doernenburg Ratio* method. This approach, published in 2015, uses the same ratios and has comparable *Diagnostic* scope. It is tabulated in Table 2-7 sans the footnotes. *Rogers Ratio* method has a *Case* for "unit normal" (**0**) whereas *IEC Ratio* method instead has a *Case* for **D1** separate from **PD**, both of which would approximate to *Rogers Ratio* method's *Case* for "low-energy density arcing – PD" (**1**). However, [1] published in 2019 still refers to *Rogers Ratio* method in its main body and not the *IEC Ratio* method, demonstrating how even established bodies may not be fully aligned regarding *Fault Identification*. Table 2 in [2, Sec. 5] provides a simplified version, shown in Table 2-8, for use in cases where the ratios do not align to any *Fault Type*.

Table 2-7: IEC Ratio Method

Case	C_2H_2 / C_2H_4	CH_4 / H_2	C_2H_4 / C_2H_6	Characteristic Fault
PD	NS[a]	< 0.1	< 0.2	Partial discharges
D1	1.0 <	0.1 – 0.5	1.0 <	Discharges of low energy
D2	0.6 – 2.5	0.1 – 1.0	2.0 <	Discharges of high energy
T1	NS[a]	NS[a] > 1.0[a]	< 1.0	Thermal fault t < 300 ºC
T2	< 0.1	1.0 <	1.0 – 4.0	Thermal fault 300 ºC < t < 700 ºC
T3	< 0.2[b]	1.0 <	4.0 <	Thermal fault t > 700 ºC

Note. a: NS = non-significant whatever the value.
b: Increasing C_2H_2 may indicate hot spot temperature is higher than 1000 ºC.

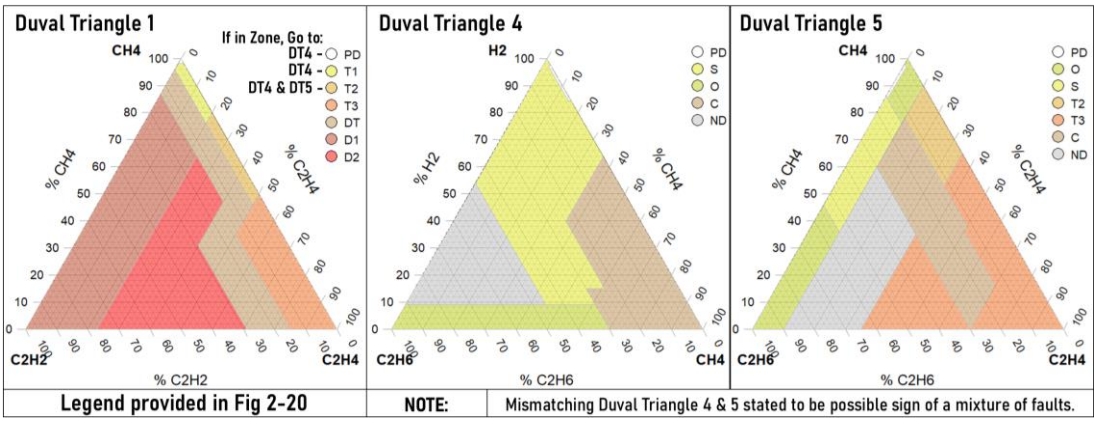
Source: Retabulated from Table 1 of [2, Sec. 5]

Table 2-8: Simplified IEC Ratio Method

Case	C_2H_2 / C_2H_4	CH_4 / H_2	C_2H_4 / C_2H_6	Characteristic Fault
PD		< 0.2	<0.2	Partial discharges
D	0.2 <			Discharges
T	0.6 > 0.2			Thermal Fault

Source: Retabulated from Table 2 of [2, Sec. 5]

One method mentioned in both [1], [2] is that first developed by Michel Duval. There are variations depending on the *Liquid Insulation*, and updated versions over time. As per [1, Sec. 6], *Duval Triangle 1* is the primary tool for mineral oil, using CH_4 for "low energy / temperature faults", C_2H_4 for "high temperature faults", and C_2H_2 for "very high temperature / energy / arcing faults". These gases are plotted on a ternary plot as shown in Fig. 2-22. Additionally, there are variations in the triangles, focussing on specific *Fault Types*. For example, *Duval Triangles 4* and *5* are stated in [1, Annex D.4] to be for obtaining more information regarding low energy or low temperature *Faults*, and high or very high temperature *Faults*, respectively. These are shown in Fig. 2-22 based on [1, Annex D.4]. Their prescribed usage was if the output from *Duval Triangle 1* was either **PD**, **T1**, or **T2**, then *Duval Triangle 4* could be referred to, if the output was either **T2** or **T3**, then *Duval Triangle 5* could be referred to, as shown in Fig. 2-22. These additional Triangles, allow for the additional categories of: **C**, **O**, and **S** as listed in Fig. 2-20. Stray gassing, **S**, is described in [1, Sec. 4] as gases generated within a TX even if no *Fault* is present and even at normal temperatures and load levels. Another formfactor of the ratios suggested by Duval are the *Duval Pentagons*, which use five gases on a single plot. This is stated to be for making interpretation easier with minimal loss in *Diagnostic* power.



Source: Redrawn from [65, Fig. 8] based on [1, Annex D.4]

Fig. 2-22: Duval Triangles 1, 4 and 5

A comparison of some of these discussed *Diagnostic Techniques* is provided in [66]. There are some well-known limitations of these ratio-based methods. For example, if there are multiple *Faults* present, or a historic *Fault*'s released gases are included with the currently active *Fault*, then the ratios may be misleading [1, Sec. 1]. Some methods such as the *Rogers Ratio* method and *IEC Ratio* method do not have a corresponding

output for all gas combinations. Even those that do, such as *Duval Triangle 1*, sometimes do not have a "healthy" equivalent output available. This then requires another method to first determine whether a *Fault* is present [1, Sec. 6], [9, Sec. 6]. However, there is no agreed upon method to determine whether a *Fault* is present.

Relative Gas Levels (Deltas)

Whether for *Fault Detection* or *Fault Identification*, it is generally encouraged to also analyse the rate of gas generation in addition to absolute gas values. As per [9, Sec. 6], "looking for trends … in several DGAs and understanding its meaning is the most important transformer diagnostic tool". Similarly, as per [1, Sec. 5]:

"…gas levels, whether high or not, in themselves are not necessarily a direct assessment of the condition of the transformer, if they are stable. A transformer with twice the gas level of another transformer, is not necessarily twice as likely to fail. On the other hand, active fault gas formation, even with low gas levels, indicates that something might be wrong…".

Gases are expected to accumulate naturally over time, so it is important to account for a TX's age [1, Sec. 5]. However, TXs are not closed systems and so the accumulated gas levels cannot be assumed as the total generated gas. Apart from interventions such as degassing that directly affects these values, gases can escape or enter, especially in free-breathing TXs [9, Sec. 6], [21, Sec. 4]. For example, [5, p. 139] states:

"A reasonably constant concentration for a readily diffusive gas such as hydrogen probably signifies a balance in the rate of the production of the gas and its loss from the system, i.e., active gas generation, whereas a constant concentration for the very soluble gas acetylene is usually taken to imply that no new gas has been produced".

Therefore, using the difference, or delta, in gas levels over a timeframe can increase the sensitivity, and focus on whether there was an indication of active gassing during that timeframe, as opposed to whether there was an indication of gassing at some point during the TX's lifespan [1, Sec. 5]. This is a recommended approach when referring to ratio methods as well. It is emphasised in [9, Sec. 6] that this is particularly useful for analysing the CO_2/CO ratio to check if there appears to be *Solid Insulation* involvement and more generally also states, "each DGA must be compared to prior DGAs to recognize trends and establish rates of gas generation". It is still essential to recognise

that different TXs may have vastly different definitions of "normal" operation depending on environment, design, and usage. As per [1, Annex G.1]:
"Attempts to assign greater significance to gas than justified by the natural variability of the generating and measuring events themselves can lead to gross errors in interpretation. However, in spite these [sic] gas-generating mechanisms are the only existing basis for the analytical rules and procedures developed in this guide. In fact, it is known that some transformers continue to operate for many years in spite of above average rates of gas generation".

It is challenging to determine a systematic approach to select the boundary points to calculate relative deltas without engineering judgement to arbitrate. Incorrect segmentation can equally lead to misleading results. [1], [2] each contain differing methods to track and compare deltas against limits, as discussed further in Chapter 3.

2.4.3. Gas Sampling

Although this thesis is not focussed on the particulars of the sampling process, it can impact DGA interpretation. DGA can be categorised in different ways. One important differentiator is how samples are obtained: whether via manual or automated sampling. A related consideration is how the sample is analysed: in a laboratory, onsite with a portable instrument, or using *Online DGA* (OLDGA). The first two approaches typically require manual sampling by a trained *Operator* whilst the third is automated once the equipment is installed.

Another categorisation is regarding the sample type. In general, DGA can be performed either directly on sampled gases, or on dissolved-in-oil gases which must first be extracted from the oil [2, Sec. 7]. The two main locations to obtain the gases directly would be via the TX headspace (or gas cushion), and within the Buchholz (or gas-collecting) relay. The method for sampling from Buchholz relays is covered in [2]. Performing DGA on gases in the Buchholz relay post-incident is recommended [1, Sec. 4.4], [2, Sec. 7], but this reactionary analysis is outside the scope of typical DGA and this thesis. There are also events that are sufficiently energetic to cause bubbling without triggering the Buchholz relay. In these cases, the bubbles may rise in the to the headspace without having time to dissolve and reach equilibrium with the gases dissolved in the oils. If these can be captured fast enough, the differences between the

gases in the headspace and in the oil can itself be of value in determining the nature and severity of the gas evolution [2, Sec. 7]. However, there also lies the potential of some of the gases in the headspace having had time to partially dissolve back into the oil, changing the relative composition that may, if taken at face-value, lead to a misleading interpretation of events.

The most common method involves sampling the oil, which contains the dissolved gases. Guidance on the sampling protocols of oil and of gases are provided in [67] and [68], respectively. There are a variety of techniques and apparatus options that could be used. An easy-to-follow methodology using glass syringes is outlined in [9, Sec. 7] for further details. The more relevant consequence is that throughout this process, there are numerous opportunities for mistakes to occur. Contamination from improper flushing or exposure to outside air, or sunlight can all significantly change the gases found. Furthermore, if the sample is not analysed on a portable device, it must be sent for laboratory analysis. The transportation process itself can introduce errors. As per [9, Sec. 7], the "sampling procedures and lab handling are usually areas that cause the most problems in getting an accurate DGA".

Actor	Action	External Factors	
		Transformer:	**Personnel:**
Transformer	▪ Provide DGA Sample	▪ Design / Size etc ▪ Other similar Assets ▪ Known defects / characteristics ▪ History / Repairs etc ▪ Refurbishments / Repairs ▪ Known defects ▪ Historic results: ▪ DGA ▪ Other test ▪ Load ▪ Current / Historic patterns ▪ Environment ▪ Temperature ▪ Current / Historic ▪ Location ▪ Indoor / Outdoor ▪ Coastal ▪ Elevation	▪ Protocols ▪ Clear Instructions ▪ Experience ▪ Follows protocols correctly ▪ Can detect improper samples ▪ Equipment ▪ Non-faulty ▪ Non-contaminated ▪ Calibrated ▪ Access to Information ▪ TX Information: ▪ Design / Size ▪ History / Repairs ▪ Load ▪ Environment
Sampler	▪ Extract DGA Sample ▪ Contain DGA Sample ▪ Label DGA Sample ▪ Package DGA Sample		
Transporter	▪ Collect DGA Sample ▪ Transport DGA Sample ▪ Deliver DGA Sample		
Analyser	▪ Receive / Store DGA Sample ▪ Analyse DGA Sample ▪ Record DGA Results		
Reporter	▪ Collate DGA Results ▪ (Collate TX History) ▪ Report DGA Records		
Interpreter	▪ Collate DGA Records ▪ Interpret DGA Records ▪ Interpret TX History ▪ Act		

Fig. 2-23: Dissolved Gas Sampling and Analysis Process

A simplified overview of the manual process is shown in Fig. 2-23. A trained *Operator* extracts a DGA sample from a given TX via a syringe. The sample must be properly contained and labelled to avoid contamination or leakage with the environment. It is packaged and sent to a laboratory for analysis. The DGA sample must be stored

appropriately in the meantime and analysed according to a specific procedure as detailed by the given laboratory. The results are then recorded. Some laboratories also provide some interpretive outputs, which may require knowledge of the TXs history. The results and interpretive outputs are recorded and handed over to an engineer responsible for the asset. They must then validate the findings, cross referencing with historic results and any contextual information regarding the specific TX, and ultimately come to a decision regarding the best action.

The gas extraction process from the sampled oil and subsequent analysis is described in [68]. The normative guidance associated with DGA interpretation is described in [2]. [68, Sec. 1] states that there are three basic methods to extract the gas from the oil for the DGA. These are:

- Extraction by vacuum (Toepler and partial degassing)
- Displacement of DG by bubbling a carrier gas through the sample (stripping)
- Partition of gases between the oil and a small volume of carrier gas (headspace).

Table 2-9: Accuracy of Laboratories using Gas Extraction Methods

Method	Average Accuracy (%)		Percentage of Inaccurate Labs	
	> 100 ppm	< 8 ppm	> 100 ppm	< 8 ppm
Partial Degassing	12	18	17	0
Stripping	19	65	60	63
Head Space	28	51	75	42
IEC Specification	15	30	–	–

Source: Retabulated from Table E.1 from [69, p. 41], original from CIGRE © 2019

A report on online DGA monitors utilising a *Round-Robin Test* (RRT) of the *Accuracy* of laboratory testing stated that "all extraction methods can provide accurate results or not, depending on how well or not they are applied" [33, Sec. 9]. However, the more recent publication revisiting the topic found that the gas extraction method can impact *Accuracy*. It found that despite the prevalence of the head space principle of gas extraction from oil in both laboratories and online gas monitors, it performed significantly worse than the partial degassing method. This is shown in Table 2-9 retabulated from Table E.1 [69, p. 41]. [70], [71] contains further information regarding the relative costs of these techniques—an important contextual factor.

Post-extraction, the gases are then generally analysed via gas chromatography, although there are other techniques, some mentioned in [21], [62], [69]. These too can

have a material impact on outputs. For example, Appendix H of [69] reports how what was once attributed to poor online monitor *Accuracy* was later discovered to be instead due to helium contamination affecting the laboratory analysis using helium as the carrier gas in a gas chromatography measurement system. In such cases hydrogen and helium spectra overlap and led to inflated values of hydrogen.

Traditionally, laboratory analyses were the primary source of DGA [33, Sec. 1] and as per [1, Sec. 5] (2019), "most DGA results fall into this category". As per Appendix 1 of [14], these were conducted approximately on a monthly-to-yearly frequency. For example, [1, Sec. 4] states that "annual DGA screening is common" or if the asset is of high significance, "online monitoring or frequent (monthly or quarterly) periodic DGA may be justified". Once an issue was suspected, the sampling frequency may have been increased to weekly or daily. [1, Sec. 4] considers frequencies of "one every few days" or faster to be *Continuous Monitoring*. However, there has been an ongoing shift towards DGA via OLCM (OLDGA) over the years as the costs decrease, scope of gases detectable increase, and the value-proposition is demonstrated via published case studies. The increased scope of detectable gases generally began in the 2000s and the market has grown considerably since 2010 onwards [33, Sec. 1], [69, Sec. 1]. This is not to imply that OLDGA is to be installed for all TXs in the foreseeable future. There is still a significant cost associated with installing, maintaining, and leveraging the data from OLDGA systems [21, Sec. 7], [33, Sec. 1], [69, Sec. 2]. The primary driver for the increased use of OLDGA is the increased sampling rate which can enable the detection of some *Faults* that would have taken longer via lab-based samples or may not been detectable at all [21, Sec. 5], [69, Sec. 2]. The increased sampling rate of OLDGA also means that the gassing rates are often more reliable [33, Sec. 1] which as previously discussed, is a metric considered a key indicator of an active problem. It is stated in [33, Sec. 1] that OLDGA is particularly useful for strategic or expensive equipment, or where significant *Faults* have already been detected. This is in line with [1], [2] which both recommend increased monitoring for TXs suspected of having issues and considering OLDGA for very high value TXs.

OLDGA systems are often recommended to be used in tandem with lab-based sampling for confirmation [2, Sec. 8] and for drift-detection purposes [21, Sec. 5]. However, oil samples taken for lab-based samples are often taken from the bottom of the *Tank* for

practical reasons whereas OLDGA monitors are often installed higher on the *Tank* or in the top oil to allow for earlier detection of gases released by the *Active Part*. However, during periods of large gassing, the two locations may have different gas compositions [21, Sec. 5], [33, Sec. 5]. This is for similar reasons as the potential differences found between gases in the headspace and in the oil previously discussed.

Neither [1], [2] have specific *Fault Detection* interpretation methods for OLDGA. Although, as per [21, Sec. 5], "with accumulation of field experience such guides [IEEE and IEC] are expected to cover on-line in addition to periodic DGA". [1, Annex A.3] states that an opportunity for improvement was to "adapt application of DGA interpretation to the use of online DGA monitors, specifically regarding the rate of change calculations". [64, Sec. 5.2] explores the topic of sampling frequency briefly and is based on [72]. [72] attempts to tie together two generally accepted concepts, one being that elevated gas concentrations and/or gassing rates can be indicative precursors to *Faults*, and that TXs more susceptible to *Faults* should be sampled more frequently. Its proposal is to scale sampling frequency based on either the gas level and/or gassing rate. This is arguably already recommended practice as elevated *Screening* outputs are often associated with increase surveillance. One aspect perhaps neglected in [72] is the finding stated in [62, Sec. C.7] that the actual limit value corresponding to a percentile can change depending on the sampling interval for gassing rates. This would potentially complicate the proposed interpolation process and introduce non-linearities. For example, [62, Sec. C.7] states that [73] found pre-failure OLDGA gassing rates at approximately ten times that of manual sampling, but that this was different for gas concentrations, hypothesising "gas concentrations are less affected by how close they are to failures". Another way to intuit this is that the gas concentration is a function of the gassing rate and time, and typically the highest gassing rates would occur the shortest duration prior to either failure or intervention.

3.Standards-Based Literature Review

Chapter Purpose

This Chapter provides a detailed Standards-driven perspective on TX CMA and the role *Uncertainty* plays within it, focussing on DGA and *Measurement Uncertainty*. *Research Theme 1A* considers the impact of changes made to the methodology in IEEE C57.104-2019 [1]. The topic is focussed on practical deployment and partly evaluated via comparisons to other candidate methodologies. To begin addressing these goals, this Chapter reviews in detail each of the methodologies, providing conceptual comparisons. *Research Theme 2* considers the impact said changes to the methodology has on *Uncertainty*. This Chapter presents a Standards-based literature review on the topic of *Uncertainty* as relating to TX CMA to conclude that there is no singular applicable methodology presented within the normative references of [1]. This provides the motivation for the contributions presented in Section 5.2.

Chapter Structure

Section 3.1 presents a detailed review and conceptual comparison of the chosen TX DGA CMA methodologies. Sub-Section 3.1.2 covers the primary focus of this thesis: IEEE C57.104-2019 [1]. Sub-Section 3.1.3 then covers its natural point of comparison: IEC 60599:2022 [2]. One addition to IEEE C57.104-2019 is an "alternative approach" detailed in its Annex F. This is the *Normalised Energy Intensity* (NEI) method, and it is covered in Sub-Section 3.1.3. Its description in [1, Annex F] is compared to its original publications, [3], [4], as well as to another industrially relevant methodology: the *Lapworth Scoring Algorithm* (LSA) [5]. The latter is covered in Sub-Section 3.1.4.

Section 3.2 presents a Standards-based literature overview of TX DGA *Screening*. Sub-Section 3.2.1 discusses the scope and role of *Measurement Uncertainty* within CMA to justify its focus in this thesis. Sub-Section 3.2.2 then discusses what information may be expected to be available for the *Screening* analysis, and how it may be used from a Standards-driven perspective. As a primary conclusion of this Chapter is that there is no clear methodology outlined in the normative references of [1], some of the limitations of the interpretation used in this thesis are highlighted and discussed. Sub-Section 3.2.3 then focusses on IEC 60567:2011 [68] as it is most relevant in establishing what information should be expected in both methodologies: [1] and [2].

3.1. DGA Interpretation Methodologies

3.1.1. Scope

The primary focus of this thesis is IEEE C57.104-2019 [1]. As the IEEE PES Transformers Standards Subcommittee aims to compare and highlight key differences between the IEEE TX Standards and the corresponding IEC Standards [74], IEC 60599 [2] is a natural point of comparison. This detailed comparison of the two methodologies is a relevant contribution, especially considering that the IEEE's Task Force's eventual output is an "internal document intended only for members of the Transformer Committee" [74]. Both these methodologies output a categorical value corresponding to recommended action(s) or interpretation of the TX's condition. In contrast, Annex F of [1] also mentions the *Normalised Energy Intensity* (NEI) method that outputs an unbound numeric transformation of the input. This thesis hypothesised that this might be more amenable for OLDGA and worth exploring. The *Lapworth Scoring Algorithm* (LSA) [5] was chosen as its point of comparison as it is a method used within industry that similarly outputs an unbound transformation of the input data.

3.1.2. IEEE C57.104-2019

Background

The IEEE methodology [1] was published in November 2019 and constituted an almost complete rewrite of the previous version published in 2008, which itself was a minor revision to the version published in 1991. At the time of the research, this was the newest active international standards from either IEEE or IEC. Although in May 2022, the IEC 60599 [2] released a newer version, replacing the 2015 version. However, this version does not alter any of the portions discussed in this thesis.

Notation

The following notation will be used henceforth for brevity:

The *Screening* outputs are given as either **L1** to **L3**, or **Ł1** to **Ł3**, depending on whether it is in reference to a single gas, or all gases combined, respectively. The tables containing the limits are given as 𝕋1 to 𝕋4 to avoid confusion with the thermal diagnoses: **T1** to **T3**. If a relevant metric is within the limit, then the table is said to "pass", otherwise to "fail".

Contextual Overview

Summarising [1] is challenging, as throughout it provides numerous suggestions for various scenarios. Nevertheless, [1] begins by stating DGA is used for:

- *Basic Risk Management*,
- *Detection and Monitoring of Abnormalities*,
- *Quality Assurance Metric*,
- *Fault Type Identification*, and
- *In-service Tripping Investigation*.

These are shown in Fig. 3-1 with the hierarchies being introduced by this thesis. Here, it is argued that *Detection and Monitoring of Abnormalities* is a means of *Basic Risk Management*. As part of this process, *Fault Type Identification*, *In-service Tripping Investigation*, and *Quality Assurance Metrics* all play a role. For example, a *Fault Type* can only be identified once detected. There is some overlap, for example, *Fault Type Identification* also uses DGA samples but in a different process not included in Fig. 3-1.

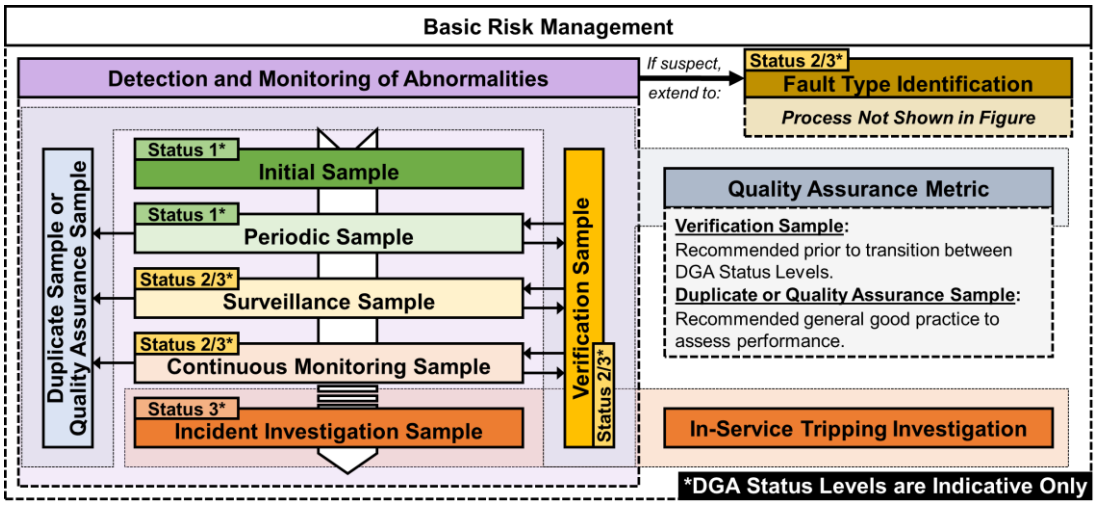


Fig. 3-1: IEEE C57.104-2019's DGA framework

In [1], the term "DGA sampling context" is used and sample types are differentiated. These too are included in Fig. 3-1 as a means of *Detecting and Monitoring Abnormalities*. Although not strictly a sequential series of events, the sampling characterisations generally imply an increased severity. The *Initial Sample* is used to help establish a *Baseline* for the TX and is not generally compared to previous samples. In [1], it refers specifically to the first sample of a reference window in time. Depending

on the outputs of the *Initial Sample*, and other factors associated with the *Consequence of Failure* (CoF) for the given TX, it is then placed in one of three *Protocols*:

- *Periodic Sampling*,
- *Surveillance Sampling*, and
- *Continuous Sampling*.

Periodic Sampling is considered the most common and applied to all TXs that have no indication of requiring further attention. Although, the higher the CoF, the lower the limit for warranting further attention. *Surveillance Sampling* is done at a higher frequency and is intended for sensitive periods of time, such as the start-up, or if a *Fault* is suspected. Once this period has subsided, the TX can be returned to *Periodic Sampling* if appropriate. If something unusual is detected, one option is opting for *Continuous Sampling* which is intended for either very high CoF TXs or ones with a suspected high *Probability of Failure* (PoF). Aside from the previously discussed benefits afforded by OLDGA, for the latter case of TXs suspected of high PoF, this *Protocol* can also help determine whether the elevated alertness is warranted, or if the TX can be returned to a more routine status. In some situations, *Continuous Sampling* may be economically unfeasible, while in others, it may be adopted directly triggered from concerning *Periodic Sampling* outputs, bypassing the *Surveillance Sampling* period. *Continuous Sampling* as defined in [1] can, but not necessarily, include OLDGA; being defined as "tested at very short time intervals (e.g., daily or several times per day)" [1, Sec. 3].

A *Verification Sample* is often recommended before transitioning from one *Protocol* to another to confirm that the triggering sample's results can be repeated and are not just a sampling *Anomaly*. Within that context, [1] often uses the term *Confirmation Sample*. A *Duplicate Sample* is very similar but is taken almost immediately whereas a *Verification Sample* is often more reactively prompted. A *Duplicate Sample* sounds similar but is instead intended to evaluate the *Repeatability* and/or *Reproducibility* of the sampling process more generally. A *Quality Assurance Sample* is an extension that can also evaluate the *Accuracy* of the sampling process as the intended outputs are known. This can be done at any point. The *Verification Sample*, the *Duplicate Sample*, and the *Quality Assurance Sample* all serve the purpose of *Quality Assurance Metric*. Lastly, should an *Incident* occur, then an *Incident Investigation Sample* is taken. This

can help determine whether the TX is likely to have been damaged and help identify the *Fault Type*. This would serve the purpose of *In-service Tripping Investigation*. Fig. 3-1 cynically implies it an inevitable state a TX will eventually experience but this is not guaranteed.

The *DGA Status* level provided as an output by [1], **Ł**, can be shown in this model to some extent, although only indicatively. The definitions of the *DGA Status* levels will be discussed further in the next Sub-Section, however they are here simplistically assumed that **Ł1** represents a TX expected to be typical, **Ł2** a TX that potentially has an issue, and **Ł3** a TX that is expected to have an issue. In this context, *Periodic Sampling* would be generally sufficient for **Ł1**, temporarily ignoring other factors such as CoF. Similarly, the *Initial Sample* would be expected to reflect **Ł1**. Conversely, a *Post-incident Sample* would be expected to show **Ł3**. Additionally, *Surveillance* and *Continuous Monitoring Sampling Protocols* are often instigated once something is suspected and are thus reflective of **Ł2–3**. The *Verification Samples* are mainly used to transition between the *DGA Status* levels and so would be most reflective of **Ł2–3**, unless they contradict the suspect sample and suggest nothing is wrong, where **Ł1** resumes. This is all caveated by the fact the these *DGA Status* levels, and how they are derived in [1] are intended in relation to the *Periodic Sampling Protocol* only. This more generalised application can nevertheless help illustrate the overall framework.

DGA Status Levels

Semantic Definition

The *DGA Status* levels are defined three times: [1, Secs 5.3, 6.1, 6.1.2]. From the phrasing, the main distinctions are the likelihood of abnormal DGA results which is in line with the function of *Detection and Monitoring of Abnormalities*. There are then secondary points regarding the *Fault Severity* mainly in [1, Sec. 6.1.2]. If the DGA outputs remain within the limits of tables 𝕋3–4 for an undefined period and *Fault Severity* is deemed low, the *DGA Status* can be dropped. Additionally, carbon oxides are considered of a lower *Fault Severity* in isolation. Another aspect is *Fault Identification*, where it states if the *Fault* identified is characteristic of a low severity type, then it is less urgent, although there is no prescribed guidance as to the practical implication.

Simplistically, the *DGA Status* levels are defined in the context of DGA outputs as:

- *DGA Status 1* (**Ł1**): probably normal,
- *DGA Status 2* (**Ł2**): possibly abnormal, and
- *DGA Status 3* (**Ł3**): probably abnormal.

However, there is an emphasis on not then assuming that the TX *State of Health* (SoH) can similarly be defined. Stated in [1] is that DGA alone cannot determine the TX SoH, and other methods and metrics should be considered in tandem. Even for DGA, [1] recommends methods for *Fault Identification* such as *Duval Triangles* once an abnormality is suspected. There is also a mention of "extreme DGA results" in [1, Sec. 6.1.2] where there is clear and urgent need for immediate additional analysis and investigation of some sorts. This is similar to the *Alarm* mentioned in [22, Sec. 10].

DGA Tables

Four tables, 𝕋1–4, are primarily used in [1] to derive the *DGA Status* level and are broadly defined in three locations: [1, Secs 6.1, 6.1.1, B.1], where the latter is focussed on 𝕋3–4. The metrics represented by 𝕋1–4 in [1, Sec. 6] are defined as:

- 𝕋1: "90th percentile gas concentrations as a function O_2 / N_2 ratio and age" in ppm,
- 𝕋2: "95th percentile gas concentrations as a function O_2 / N_2 ratio and age" in ppm,
- 𝕋3: "95th percentile values for absolute level change between successive laboratory DGA samples" in ppm, and
- 𝕋4: "95th percentile values from multi-points (3-6 points) rate analysis of laboratory DGA samples with all gas levels below Table 1 values" in ppm.

These tables are in Annex A but should be considered in context of the entirety of the guidance given in [1]. 𝕋1–2 can be considered a pair as they are the 90th and 95th percentile of absolute gas levels [ppm], respectively. This is conceptually very simple and would highlight TXs with unusually high accumulated gas levels relative to the dataset. This would be most relevant during the *Initial Sample* where gas rates cannot be calculated. For [1] however, they are used for all *Protocols* with equal emphasis. 𝕋1–2 have specific limits for the gases according to TX type (or O_2/N_2 ratio) and TX age.

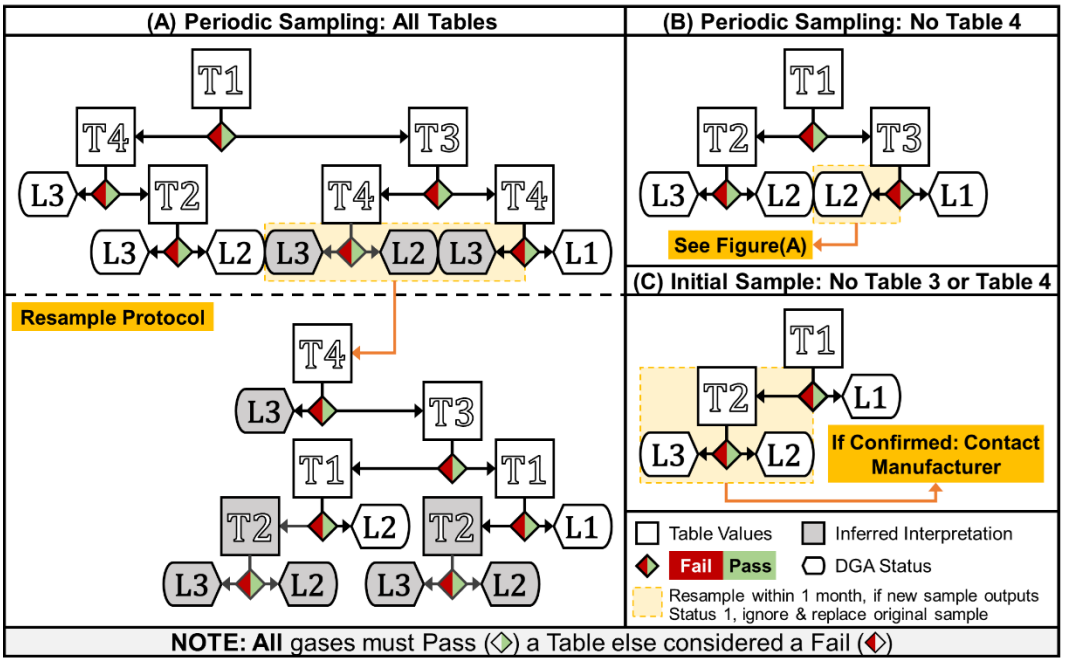
𝕋3–4 are more unusual in their implementation. Both consider the change in gas levels at the 95th percentile, but 𝕋3 considers only the difference, or delta, between consecutive samples [Δppm] whereas 𝕋4 uses the rate of change in gas levels

normalised to a per year basis [ppm/year] calculated using multiple samples. 𝕋4 uses 3-6 samples over 4-24 months to calculate a linear regression line, from which the slope normalised to a one-year duration is used as the metric. As per [1, Sec. 6], "if more than 6 data points are available, use the six most recent data points, not exceeding two years, to compute the rates". 𝕋3–4 have specific limits for the gases according to TX type. 𝕋4 also differentiates based on the overall interval between the first and last sample selected for the rate calculation by using two categories: 4-9 months and 10-24 months. In contrast to 𝕋1–2, there are no categories for TX age.

Derivational Definition

The other aspect of the *DGA Status* levels to discuss are their derivations which are primarily derived via the use of 𝕋1–4. 𝕋1–4 are applied to each gas separately before being combined. The gases included are: H_2, CH_4, C_2H_6, C_2H_4, C_2H_2, CO, and CO_2. It is challenging to formalise the derivation of the *DGA Status* levels comprehensively given the amount of discretion [1] advises regarding the classification. There are numerous edge cases where it suggests overriding the *DGA Status* based on various factors. The original work for the automated implementation related to the thesis was published in [75]. Fig. 3-2 is modified from [75, Fig. 1] to include the derivation for the cases where 𝕋3–4 may be incalculable, as based on [1]. This derivation process is repeated for every available gas. The worst case amongst all available gas outputs, **L**, is selected to then determine the overall *DGA Status* level, **Ł**, of the given sample.

A complicating factor of the derivation process is the inclusion of the *Confirmation Sample* as this can retroactively impact the validity of prior samples depending on its output. There is also some confusing guidance regarding this *Confirmation Sample*. Step 4c from [1, Sec. 6] states to "compute rates with the confirmation sample replacing the previous value" whilst Step 5 states that "a confirmation sample … will allow the computation of the rates (e.g., 3 samples in 2 years)" for cases where there are originally only two samples. This thesis assumes this is done only for the case where there are otherwise insufficient samples. Some parts of the process outlined in [1, Sec. 6] were not explicit, and so were assumed based on context and highlighted as such in Fig. 3-2.



Source: Modified from [75, Fig. 1] based on interpretation of [1]

Fig. 3-2: Interpreted IEEE C57.104-2019's DGA Status derivation

Fault Identification

There are no significant contributions to TX *Fault Identification* in [1]. The main body includes *Rogers Ratio* method and *Duval Triangle 1*. The Annexes include the *Key Gas* method, *Doernenburg Ratios* method, *Duval Pentagon 1*, and *Duval Triangles 4* and *5*. However, the first two methods are discouraged as outdated or unreliable.

It is generally recommended to only apply *Fault Identification* techniques once a *Fault* is suspected. As per [1, Sec. 5], "when there is an indication of a problem, a fault identification or diagnosis should be obtained by a reliable technique". The specific criteria for this are somewhat ambiguous. For example, [1, Sec. 5] states that "if analysis of a test result produces a DGA status of 2, or if any unusual shift in gas pattern suggests an anomaly, a fault diagnosis should be obtained". It is not explicit in what constitutes an unusual shift in gas pattern. Additionally, gas ratios are often used to detect shifts in patterns but are here instead reserved for *Fault Identification*.

Annexes D.8 and D.9 in [1] also contain information regarding interpretation. Annex D.8 is focussed on the carbon oxides gases (CO and CO_2). It states that historically, these gases were considered good indicators of paper involvement in *Faults* but that some recent investigations indicate that this is not always true. The carbon oxide ratio is emphasised: as per [69, Sec. 2], "very few if any cases of faults in paper have been

reported which could reliably be detected by significant increases if carbon monoxide alone, without knowledge of carbon dioxide values and/or of the other gases".

According to [1], if CO concentrations are high whilst the CO_2/CO ratio is low, then the degree of hydrocarbon gases should be checked. If they are not rising in significant amounts, then this is "NOT an indication of a fault in paper, particularly in closed transformer, but are rather due to mineral oil oxidation under conditions of limited supply of O_2" [1, Annex D.8]. However, if they are rising in significant amounts, then this may be an indication of a *Fault* in paper but should be also confirmed via other methods. If CO_2 concentrations are high whilst the CO_2/CO ratio is also high, then ideally the furans would be checked to confirm if they are also high. This would then be "an indication of the slow degradation of paper at relatively low temperatures (<140 °C) ..." [1, Annex D.8].

If the carbon oxides are below T1 limits, then they are said to correspond to "normal gassing in transformers without faults". Although it proceeds to state that "zero or very low rates of change of CO and CO_2 do not necessarily mean the absence of a fault in paper" [1, Annex D.8]. It is being assumed that the first quote is intended to mean that carbon oxides below the limits do not themselves indicate paper-related *Faults* but neither do they necessarily rule out the possibility. In other words, the previous paragraph's definition of 'high' is taken as to exceed said limits.

Annex D.9 has two short notes on the O_2/N_2 ratio, and the C_2H_2/H_2 ratio. The first is that if the O_2/N_2 ratio is decreasing, then it indicates overheating and oxidation of mineral oil and it can be used as supporting evidence of thermal *Faults*. If the O_2/N_2 ratio is instead increasing, then it may "indicate leaks in the air preservation system of transformers (membrane or nitrogen blanket)" [1, Annex D.9]. The second note is that if the C_2H_2/H_2 ratio is greater than three, it may "indicate leaks or contamination from the tap-changer compartment into the main tank" [1, Annex D.9].

Uncertainty

Section 2.3 included a general overview of the topic of *Uncertainty* and a more detailed literature review will be provided in Section 3.2, including the content in [1]. This Sub-Section is focussed more specifically on the actionable components of the content in [1]. In general, [1, Sec. 1] warns of three sources of uncertainty:

- Normal variation due to load and environmental conditions,
- Unavoidable random "noise" from *Measurement Uncertainty*, and
- *Data Quality* issues.

These can be considered as ordered somewhat by difficulty to control. *Data Quality* issues would ideally be avoided by simply not making *Blunders*, or gross errors unlikely to be repeated nor reliably predicted. In contrast, *Measurement Uncertainty* may be reduced by better procedures but cannot be fully eliminated. Normal variations are intrinsic to measurement and reflect the true nature of the TX that would be hard if not impossible to remove. The two main practical tools suggested for mitigating the risk of false identification are *Verification Samples* and *Confirmation Samples*. The first is intended to validate the accuracy of overall sampling process. The latter is used to reduce the likelihood of the readings being just an anomalous sample that cannot be repeated. As per [1, Sec. 5]:
"It is important to understand that a review of data cannot identify all possible data quality problems. Therefore, when unexpected or alarming results are obtained, it is highly advisable to collect and process another sample to confirm results".

Data Quality

From the listed potential sources of errors impacting data quality in [1, Sec. 5], the ones that have actionable suggestions are highlighted here.

Hydrogen Levels

It is stated in [1, Sec. 5.1.4] that if H_2 drop significantly whilst the O_2/N_2 ratio is ~0.2 for sealed TXs, then it is an indication of air exposure. Similarly, if the H_2 drops significantly whilst the O_2 and N_2 levels are at about saturations values, then this also is an indication of air exposure. Additionally, it states that if H_2 levels are "always extremely low, even when other combustible gases are not, and especially when hydrogen is also chronically low in other transformers, there may be a problem with sampling technique, leaky syringes, or measurement" [1, Sec. 5.1.7]. It states that a *Verification Sample* may be useful to confirm that H_2 is being measured correctly.

Inconsistent Values

It is stated in [1, Sec. 5.1.6] that "if there are large inconsistencies in several successive samples, there could be a sampling or measurement problem". Regarding the O_2/N_2

ratio, it states that if it were to rise rapidly whilst H_2 decreases then there may be an air exposure problem. If the ratio begins to near 0.4 or 0.5 for sealed transformers, then to check for air leaks as this is near the saturation value for air dissolved in mineral oil. Rapidly increasing O_2 levels may indicate a damaged bladder in the *Conservator* where applicable. For open breather TXs, the total O_2 and N_2 levels could increase above saturation values in case of air ingress into the sample. [1, Sec. 5.1.8]

DGA Reliability

Accuracy and consistency are highlighted as two key considerations in [1, Sec. 5.2].

Accuracy

It defers to [33], [68] for the topic of accuracy, stating that ±15% *Accuracy* for DGA results is recommended. It suggests the use of a *Verification Sample* to ensure the *Accuracy* of the laboratory analysis. Also stated in [1, Sec. 5.2.1] is that at concentration levels below approximately five times the detection limit, *Fault Type Identification* is unreliable due to the high relative *Measurement Uncertainty*. It explicitly states: "it is not recommended to attempt fault identification using the methods described … if all of the gas levels are below that Table 1 values". C_2H_2 is highlighted as being especially problematic as it is often a low value whilst at the same time being very influential in the output of most *Fault Type Identification* methods. It cautions that C_2H_2 levels even above 𝕋1 limits can still be below the aforementioned five times detection limit.

Consistency

Regarding consistency, [1, Sec. 5.2.2] states that "when DGA results consistently fluctuate widely (30% or two or three times the values in Table 3) from one sample to the next, it usually indicates sampling or analytical errors" and suggests they should not be relied upon until the cause is understood: recommending resampling. A change in circumstances could be a cause. A *Verification Sample* could also be considered to check whether there is an issue with the sampling process.

Continuous Monitoring

For *Continuous Monitoring* (which is not necessarily OLDGA), [1, Sec. 5.3.4] states that rates should be treated differently and that the suggested screening norms no longer apply. It states that often higher rate values are used but that "each situation is unique".

It also states that "care should be taken to account for the intrinsic fluctuations of the DGA levels generated by the monitoring process".

DGA Table Values

Although [1, Sec. 6] provides default values within 𝕋1–4 for those with insufficient data to derive their own, it does not include the underlying distributions used. This means that any deviation from the default values has an unknown implication regarding the shift in the percentiles of the original dataset. There is perhaps some limited scope to assume a distribution shape and use 𝕋1's 90th % and 𝕋2's 95th % to estimate a distribution. However, firstly, the values in the tables were subject to an unconventional rounding schema as described in [1, p. 45], and they were also combined with adjacent values when they differed by <35%. Secondly, this would not help with estimating 𝕋3–4 values. Lastly, Figure A.8 in [1] shows the 90th % value for CO for O_2/N_2 ratios above and below 0.2 across the different component datasets that were eventually aggregated. It shows that there is a large variance, and that even the relative values between the two O_2/N_2 ratios varied significantly across the datasets.

As an example of a value in 𝕋1–4 that may be undesirable, the 𝕋2 limit for CH_4 decreases from 60 ppm to 30 ppm as the TX age exceeds 30 years. It is not clear why or if it should be considered a generalisable result applicable elsewhere. Given that this is counter to all other gases, one may wish to adjust this limit. However, without an applicable distribution, it is challenging to justify any particular change.

C_2H_2 seems an outlier and potentially problematic due to having a limit of zero for both 𝕋3–4. Put simply, if the gas value is fixed about a mean with any degree of noise, then every fluctuation due to said noise would cause the one or both 𝕋3–4 to fail and thus **Ł2–3** depending on the circumstances. Given that the limits for 𝕋1–2 range from 1–7 ppm, this would imply a non-trivial number of occasions where the gas values are expected to be above the conventional *Limit of Detection* (LoD) of 1 ppm, but below these values, and thus potentially affected by spurious flagging due to noise. Especially for OLDGA, where rounding manual rounding is typically not done, a limit of zero is untenable. Although it is reiterated that [1] is not intended for OLDGA.

3.1.3. IEC 60599:2022

Background

The IEC methodology [76] was published in 2015 and updated with minor revisions in the superseding 2022 version [2]. The document [2] will be referred to as IEC, and [1] as IEEE. A challenge with its implementation compared to [1] is the absence of default values for some limits, requiring other sources and introducing further subjectivity to this comparison. This Sub-Section compares the IEC methodology to the IEEE where possible, but it is impractical to trace everything to its respective sources for further analysis or interpretation. As a result, some differences highlighted will remain as unresolved comments.

IEC Outputs

Semantic Definition

The methodology outlined in [2] can be considered as a two-step process: *Screening* that potentially leads onto a *Fault Identification* process. Its primary *Screening* output would be either a *Typical* condition, *Alert* condition, or *Alarm* condition. These outputs do not directly correspond to *DGA Status 1–3* based on the descriptions in [1].

It is stated in [2, App A.2.3] that "any gas formation below typical values of concentration and rates of gas should not be considered an indication of "fault", but rather as "normal gas formation"". When compared to the description for **Ł1**, which states these TXs "are considered probably normal, per DGA results statistics" [1, Sec. 6.1], it could be argued the **Ł1** aligns well with IEC's *Typical*. However, [1, Sec. 6.1] states that **Ł1** represents "low gas levels and no indication of gassing". If taken at face value, all limits in 𝕋4 of IEEE should be 0 ppm/year, which they are not. Therefore, this is interpreted as loose wording that allows natural variation.

Stated in [1, Sec. 5.3] is that for **Ł2**, one should resample and monitor possible gas evolution, and to perhaps consider OLDGA. This is clarified in [1, Sec. 6.1] where it states if the *Diagnosis* indicates a more serious issue than say **PD**, then "increased sampling frequency should be maintained or started". This is similar to the default recommended action for an *Alert* condition, stating to "institute more frequent sampling" and to "consider on-line monitoring" [2, Fig. 1].

Ł3 is characterised in [1, Sec. 6.1] as "high gas levels and/or probable active gassing" and "probably suspicious". It recommends considering OLDGA, increased sampling, and obtaining expert opinion. The IEC in default recommends "immediate action" for an *Alarm* state, including to "consider on-line monitoring, inspection or repair". This is more stringent than **Ł3** and more aligned with the IEEE's *Extreme DGA Results*: "immediate investigation and operating restrictions should be initiated" [1, Sec. 6.1]. *Alarm* concentration values are defined in [2, Sec. 8.2] as the point from which the PoF is "sufficiently high to require urgent competent decisions and/or actions".

As another comparison, Appendix C.8 in [62] referencing [2], suggested four levels:

- Typical gas levels: "increase oil sampling intervals from yearly to monthly, consider performing complementary tests … or reduce loading".
- Intermediate 1 levels: "increase oil sampling to weekly".
- Intermediate 2 levels: "consider installation of on-line multi-gas monitors and inspection depending on results of complementary tests".
- Pre-failure (alarm gas levels): "consider immediate action, removal from service for repair or replacement depending on damage observed".

Splitting *Alert* into *Intermediate 1* and *Intermediate 2* allows for better alignment with **Ł1–3**, the *Alarm* level would then correspond to IEEE's *Extreme DGA Results*.

One other difference between [1] and [2] is that the latter suggests a distinct method for *Fault Identification* whereas the former only refers to other established methods.

DGA Tables

The overall methodology relies on two tables, one for gas concentration and one for gassing rates. Both tables have two limits, termed *Typical* and *Alarm* levels. Here, they are termed L_1, L_2, G_1, and G_2 for the first and second limits for gas concentration (L) and gassing rates (G), respectively.

The overall premise is stated in [2, Sec. 8.1] to be based on a reference to [64, Sec. 5]; that PoF is related to gas concentrations and rates of gas formation. In [2, Sec. 8.1], it is stated that "below certain concentration levels (quotes as typical or normal), the probability of having a failure is low", citing [77] saying that it is "typically 10 % according to CIGRE when using oil sampling". It should be noted that the correct

interpretation is that of the TXs that have a failure-related event in service, 10% did so whilst in this range. This is stated more clearly in [77, p. 296] and visualised in [77, Fig. 1]. Appendix C.5 of [62] expands on this. The generic annual failure rate is stated in [77, p. 296] as 0.3%, which is much more in line with other literature.

Whilst within the *Typical* limits, [2, Sec. 8.1] states that the equipment can be considered healthy and improbable to fail and therefore concludes that a "first rough screening between healthy and suspect analyses" can be obtained this way. PoF is stated in [2, Sec. 8.1] to potentially increase significantly at values much above these typical values. this is illustrated on a per-gas basis in [62, Fig. C.1]. Above these values, [2, Sec. 8.1] states the situation is then considered critical, for even though there may not be a failure, the risk is high. Regarding the rates of gas increase, [2, Sec. 8.3] states if there is no change, then likely the *Fault* has disappeared, or the gassing rate is similar to the rate at which that gases are leaking (mainly for air-breathing).

To determine the two limits, [2, Sec. 8.1] suggests that utilities with sufficient DGA data can generate a *Cumulative Distribution Function* (CDF) stratified by relative TX properties to examine the PoF at specific gas concentration levels. As per [2, App. A.1], "individual networks are strongly encouraged to calculate the typical values for their own population". Although, [2, Sec. 8.1] notes that this requires a lot of data. It is stated in [2] that ultimately, limits should be decided by the user / manufacturer / experts.

The topic of norm selection is discussed in [1, Sec. 5.4] where it states that twice the value does not necessarily mean twice the PoF but does not provide specific guidance regarding implementation. It is stated in [1] that insulating liquid volume, TX rating and voltage did not produce significant differences, which differs slightly from [2]. Furthermore, [2, Sec. 6.1] states that "typical values in open and closed transformers are relatively similar" whereas [1, Sec. 5.4] states that "the ratio of O_2/N_2 ... have a large influence on the typical levels of gases". The latter is also corroborated by [4, Sec. B]: "transformers with more oxygen in oil tend to have lower concentrations of dissolved combustible gases", explaining sealed TXs "generally have very low oxygen content".

If no adequate dataset is available, [2, Sec. 8.2] states users may use values observed on other networks. [2, App. A.2] provides a range of values for *Typical* limits based on worldwide observations made by the IEC and CIGRE. It notes that values are

inapplicable to TXs that are frequently degassed. Table A.2 and Table A.3 in [2, App. A.2] provides the range of 90th % values found in their dataset which is most relevant to the *Typical* limit for the gas concentrations and gassing rates, respectively. These values were from [77] published in 2006 with the gas concentrations based on "about 25 electrical networks worldwide and including more than 20 000 transformers" and the gassing rates based on "4 electrical networks including more than 20k DGA results".

However, [2] does not provide values for the *Alarm* limits. Furthermore, its source [77] also does not provide *Alarm* limits for the gassing rate. This thesis considers [62] from the same body and similar authors to be an update to [54] and is used instead as it contains the necessary limits. Therefore, Table C.7 and C.8 from [62] are used instead. Appendix C.8 of [62] presents four limits as discussed earlier. This makes it challenging to select a limit to represent *Alarm* levels. It is stated in [77, Sec. 2.5] that:
"Alarm gas concentration (AGC) values are defined as values intermediate between the typical values (below which the transformer is considered as relatively safe), and the pre-failure values (above which a failure may be imminent). The choice of alarm values is dependent on the tolerance to risk of maintenance personnel, also on economic and strategic considerations (cost of increased monitoring), so it is left to users to decide."

Therefore, in line with this, this thesis takes L_1 / G_1 as the *Typical* quantity, and L_2 / G_2 as the *Intermediate 2* quantity. Table 3-1 tabulates the relevant gas concentration level limits from Table C.7 from [62].

Table 3-1: Limits for Gas Concentration Levels used for IEC Implementation

Limit Categories		**Gas Concentration Levels** [PPM]						
Case*	Level	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO	CO_2
L_1	Typical	118	85	111	56	5	700	6300
	Intermediate 1	200	135	210	120	19	970	11600
L_2	Intermediate 2	280	180	300	200	40	1180	16700
	Pre-failure	725	400	900	800	450	2100	50000

*: L_1 is Alert limit, L2 is Alarm limit.

Source: Values from Table C.7 from [62], original from CIGRE © 2019

Table 3-2: Limits for Gas Rate Levels used for IEC Implementation

Limit Categories		**Gas Rate Levels** [PPM / year]						
Case*	Level	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO	CO_2
G_1	Typical	21.6	15.6	14.4	12	0	192	1620
	Intermediate 1	72	60	84	48	0	720	6240
G_2	Intermediate 2	132	144	204	132	12	1560	13680
	Pre-failure	1080	1800	3960	1800	180	17040	150000

*: G_1 is Alert limit, G_2 is Alarm limit.

Source: Values from Table C.8 from [62], original from CIGRE © 2019

Following similar reasoning as stated for the concentration levels, this thesis uses the *Typical* and *Intermediate 2* values from Table C.8 from [62]. Table 3-2 presents the results on a per year basis whereas the original was a per month basis. This transformation was applied here to be the same units as in the equivalent table in the original IEC guidelines as well as to compare the values later to the IEEE guidelines which also uses a per year basis as its limit. This is interpreted as a similar action to how [2, App. A.2] states its table can be converted from ml/year to ml/day when TX oil volume is known. However, care should be taken when making such transformations, as rescaling can impact results. It is stated in [1, Annex B] that "percentile values of rates computed with a certain time interval group (e.g., yearly DGA) are not the same for a difference interval group (e.g., quarterly DGA). The difference is quite large and grows exponentially as a function of the inverse of time difference between DGA results…". This would therefore likely mean the 'true' value for the per year basis would be lower.

DGA Tables Comparison

The L_1 representing the 90th % of the absolute gas value should be same in scope as $\mathbb{T}1$ of the IEEE. Similarly, L_2 should be somewhat comparable to $\mathbb{T}2$, though likely higher given the previous discussion regarding relative severity. However, the gassing rate for the IEC as a metric, is somewhere in between $\mathbb{T}3$–4 in the IEEE. It is the rate calculated between samples, which is like the difference between samples of $\mathbb{T}3$ in the IEEE, but it is normalised by the interval, making it a rate comparable to $\mathbb{T}4$. Another difference between it and $\mathbb{T}4$ is that it is using only 2 points and not 3–6 points.

For a very approximate comparison of these bounds, the top plot of Fig. 3-3 shows the L_1 and L_2 limits of the IEC for the absolute gas values as well as the $\mathbb{T}1$–2 limits of the IEEE. The bottom plot shows the G_1 and G_2 limits of the IEC for the gassing rate as well as the $\mathbb{T}3$–4 limits. $\mathbb{T}1$–4 have several categories to select from depending on factors such as age and O_2/N_2 ratio. The latter here is termed the air ratio, with an air ratio $\leq$ 0.2 being termed *Sealed TX* and above as a *Free-breathing TX*. To select a suitable limit for $\mathbb{T}1$–4, the average value for the *Sealed TX* and the average value for the *Free-breathing TX* was averaged with a two-to-one weighting given to the former. This is based on [1, Fig. A.5] which shows approximately twice as many DGA samples obtained

for *Sealed TXs*. It may be possible to utilise [1, Fig. A.2] and [1, Fig. A.4] to also weight based on age categories in 𝕋1–4, but this was not pursued here.

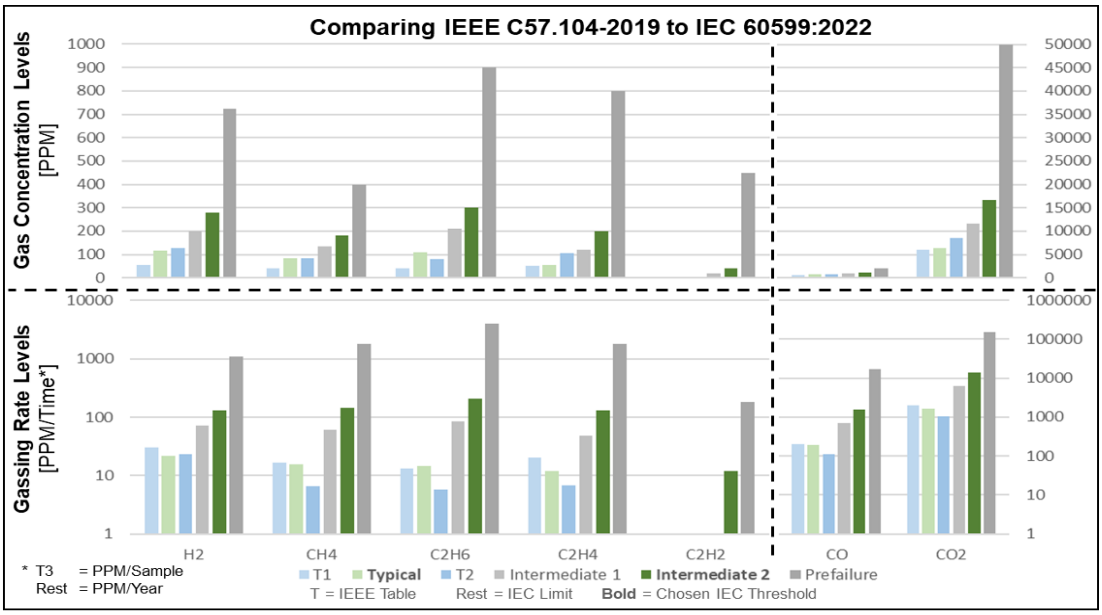


Fig. 3-3: Comparing Limits between IEEE C57.104-2019's and IEC 60599:2022

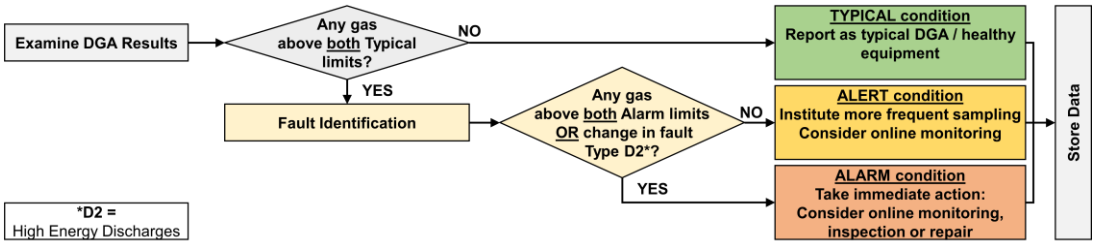
There is no clear alignment of L_1 of the IEC with either 𝕋1–2 of the IEEE visible in Fig. 3-3, but rather a mixture of both. Arguably, 𝕋1 seems more demanding and 𝕋2 broadly more like L_1. *Intermediate 1*, *Intermediate 2*, and especially *Pre-failure* values of the IEC are all higher than 𝕋2. 𝕋3 is arguably better aligned with G_1 of the gassing rate, though it is again noted these not directly comparable metrics. This is nevertheless interesting as it may suggest that gassing rate for G_1 may be simply measurement noise as that was the stated *Screening* purpose of 𝕋3. There is also no direct escalation between 𝕋3–4 as there is between G_1 and G_2 of the IEC. 𝕋4 uses linear regression trying to reduce *Uncertainty* is more onerous than the others, presumably since it can afford to discount measurement noise a limited extent. It seems reasonable to consider L_2 and G_2 to be at a much higher level of severity and is therefore consistent with the interpretation that the *Alarm* output is to be treated as a higher level of severity than **Ł3**.

<u>Derivational Definition</u>

A simplified flowchart of the process outlined in [2, Sec. 9] is shown in Fig. 3-4 adapted from [2, Fig. 1]. The derivation process at first seems simpler than that of the IEEE approach but appears underspecified upon closer inspection. It states that if all gas levels **and** gassing rates are below *Typical* values, then it can be reported as "normal DGA/healthy equipment", but if any gas's level **and** rate is above *Typical* values, to then

proceed to calculate the ratios and identify the *Fault* using Table 2-7 or Table 2-8 based on Table 1 and Table 2 from [2, Sec. 5.4], respectively. The same phrasing is repeated both in [2, Fig. 1] and in Clause A) of [2, Sec. 9]. It is therefore not clear on the recommendation should a gas level be above the *Typical* value whilst its gassing rate is below. If all gases are below the *Alarm* limits and the *Fault Identification* does not indicate **D2**, then it recommends considering more frequent sampling and potentially online monitoring. Otherwise, it is an *Alarm* condition and immediate action is recommended: either online monitoring, inspection, or repair.

Given that there is not the same problematic phrasing for the *Alarm* condition, this thesis will assume that interpretation: i.e., if **any** gas level **and** gassing rate are above *Typical* values, then it can be assumed **not** a *Typical* condition. As a result, this becomes a major difference between [2] and [1] as the latter can escalate its output due to either concentration levels or gassing rates. This seems counter-intuitive as if there is severe gassing identified, then perhaps action ought to be considered regardless of historic gassing (i.e., its accumulated gas concentration). This therefore remains a potential issue. However, this may be partly mitigated by having gases associated with the high-risk **D2** able to escalate the *Screening* unilaterally.



Source: Adapted from [2, Fig. 1]

Fig. 3-4: Simplified IEC 60599:2022 Process

Fault Identification

A general background is first provided by [2] regarding the mechanisms of gas formations and how they can be used to identify *Faults* based on their values and compositions. This can be compared to [1, Secs 4.0, 4.1, 6.2, Annex C] and is broadly similar in scope and content. Three primary gas ratios are highlighted in [2], C_2H_4/C_2H_6, CH_4/H_2, and C_2H_2/C_2H_4, that it later uses for *Fault Identification* of a list of visually detectable *Fault Types*.

Although there is some overlap between **D1** and **D2** where dual attribution should be given, [2, Sec. 5] highlights that the distinction is kept as **D2** can be significantly more damaging and potentially necessitate different preventive measures. As shown in Fig. 3-4, **D2** can escalate the *Screening* output unilaterally whereas **D1** cannot. In cases where ratios fall outside the range, [2, Sec. 5] states it can be considered that there is a mixture of *Faults* or that a new *Fault* is superimposed onto high background gas levels. In these cases, [2, Sec. 5] states that its Table 1, shown in Table 2-7, should not be used, but instead to look at the graphical representation to see which it is closer too. Alternatively, the simplified Table 2, shown in Table 2-8, can also be used to get a rough indication of *Diagnosis*.

When discussing the evolution of *Faults*, [2, Sec. 5.9] states that they often start as incipient *Faults* of low energy that escalate and it is therefore important to not only consider the increase in gas concentrations, but also a potential shift / evolution in the *Fault Type* over time. It is stated in [2, Sec. 6.1] that, if necessary, the difference between subsequent values can be used to calculate ratios if the new *Fault Type* seems different than from the previous analysis. The topic of "mixtures of faults" was also discussed in [1, Annex D.6] where it similarly suggests using the change in values. To track the evolution of *Fault Type*, it recommends looking at the ratios graphically and highlights the usefulness of *Duval Triangle* in always providing a *Diagnosis*. It is acknowledged in [2, Sec. 5.9] that other times, "instant final breakdown can occur without warning".

Aside from the three ratios used for the primary *Fault Identification*, three others are discussed: CO_2/CO, O_2/N_2, and C_2H_2/H_2.

<u>CO_2/CO</u>

There is a detailed section on the use of CO_2/CO which can be compared to [1, Annex D.8] which itself refers to the IEC—though it differs in some of its guidance. Based on the ordering of the content, it is assumed that the IEEE was similarly inspired by the IEC content / sources for the O_2/N_2 ratio and C_2H_2/H_2 ratio, [1, Annex D.9].

The formation of carbon oxides is stated to increase rapidly with temperature in [2, Sec. 5.5], and that can be used to indicate a *Fault* involving paper. Although, [2, Sec. 5.5] warns that "some localised faults in paper do not produce significant amount of CO and CO2 (or furanic compounds)", and that carbon oxides should not be used in isolation

to determine paper involvement of *Fault*. Rather, supporting evidence such as gas formation should be used. This aligns with guidance in [1, Annex D.8].

It is stated in [2, Sec. 5.5] that high values of CO (giving 1,000 ppm as an example) and ratios of $CO_2/CO < 3$ are generally considered an indication of paper involvement in a *Fault*, with possible carbonisation, if in the presence of other gases. This aligns with [1, Annex D.8]. It is highlighted in [2, Sec. 5.5] that some recent TXs that have low atmospheric mixing can have CO accumulation without irregularities and as such, the presence of other gases is an important criterion. It is similarly stated in [64, Sec. 5.7] that sealed TXs (and thus low atmospheric mixing) can have ratios of $CO_2/CO < 3$ without a *Fault*.

High values of CO_2 (giving 10,000 ppm as an example) and ratios of $CO_2/CO > 10$ are stated in [2, Sec. 5.5] to be generally considered an indication of mild overheating or oil oxidation, especially in free-breathing TXs. It explains that CO_2 can accumulate more rapidly than CO in free-breathing TXs operating at changing loads as their solubilities in oil differ, and that combined with long-term degradation of paper at low temperatures, this can lead to high CO_2/CO ratios in aged equipment. This differs slightly from [1, Annex D.8] which recommends higher ratios of $CO_2/CO > 20$ and states the mild overheating to be < 160 °C as opposed to [2, Sec. 5.5]'s < 140 °C. The referenced justification from [64, Sec. 5.7] refers to [77, Sec. 4], stating that: "it has been previously shown by TF11 ... that CO_2/CO ratios of 20 to 50 are formed when overheating prototype at 160 to 130°C, respectively". This appears consistent with Table 19 from [77, Sec. 4], showing a decrease in CO_2/CO ratio as paper temperatures increased. If following this assumption, then the guidance from [2, Sec. 5.5] is more conservative than [1, Annex D.8].

One aspect not as emphasised in [1], is regarding the use of differences to obtain reliable CO_2/CO ratios by accounting for possible absorption of atmospheric CO_2, as well as background carbon oxides accumulated over time. Indicatively, [2, Sec. 5.5] states that an air-breathing equipment saturated with approximately 9–10% of dissolved air can contain up to 300 ppm atmospheric CO_2.

$\underline{O_2/N_2}$

It is stated in [2, Sec. 5.6] that at equilibrium, free-breathing TXs have approximately 32,000 ppm and 64,000 ppm for O_2 and N_2, respectively. It states that in service, the O_2/N_2 ratio can decrease due to oil oxidation and/or paper ageing; clarifying that this is dependent on the relative rate of O_2 consumption as compared to its replenishment via diffusion. It also states that factors such as load and preservation system be impactful, but in general, a ratio of $O_2/N_2 < 0.3$ is generally considered indicative of excessive consumption of O_2. Although, it does not explain the implication. [1, Annex D.9.1] clarifies that the decreasing ratio can potentially be used to confirm thermal *Faults*, although it does not provide any numerical values regarding ratios. Furthermore, it states that "increasing values may indicate leaks in the air preservation system of transformers (membrane or nitrogen)".

$\underline{C_2H_2/H_2}$

It is stated in [2, Sec. 5.7] that OLTC contamination can be mistaken for **D1** *Fault Type*, as indicated by a ratio of $C_2H_2/H_2 > 2$ or 3. Similar is stated in [1, Annex D.9.2] for a ratio of $C_2H_2/H_2 > 3$. This is noted in [2, Sec. 5.7] to be of lesser concern with modern OLTCs which are designed not to contaminate.

Uncertainty

Uncertainty is discussed in [2, Sec. 6], which states that sampling and analysis should follow [67] and [68]. Based on these, it states values below the analytical limit should not be recorded numerically, and that, if samples taken over a short period of time, such as days or week, show inconsistent variations, they should be eliminated or corrected. On this point, [1, Sec. 5.2] adds illustrative values of "30% or two to three times the values in Table 3" between samples. [2, Sec. 6] then states if the "gas ratios are different from those for the previous analysis", then a new *Fault* may have developed. It is not clear the extent the ratio should change prior to accepting this assumption. In such cases, it recommends calculating the ratios based on the changes between subsequent samples, and highlights this is particularly true for carbon oxides.

Lastly in [2, Sec. 6], it notes in reference to [68], that above ten times the *Analytical Detection Limit* (S), the *Uncertainty* is typically ±15% on DGA values, rising to typically ±30% at five times S. It states care should be taken interpreting gas ratios at gas values

below ten times S. Similar guidance and the same source, [68], is referred to in [1, Sec. 5.2], although it goes further by stating "it is not recommended to base fault type identification or practical decisions on such low values [below about five times LoD] without some confirmation of their accuracy". This thesis generally uses the term *Limit of Detection* (LoD) in lieu of *Analytical Detection Limit* and reserves S for its mathematical notation.

Calculating the rate of gas increase since the last analysis, "taking into account the precision on DGA results" is recommended in [2, Sec. 9]. Similarly, [2. App. A.2.5] states that "when calculating typical rates, intervals should be chosen to provide an acceptable accuracy or results". However, the practical interpretation is not elaborated.

Continuous Monitoring

OLDGA is stated in [2, Sec. 8.5] to be "particularly well-suited for detecting non-typical rates of gas increase occurring within minutes, hours, or weeks, which is generally not possible with routine oil samplings done at monthly or year intervals" citing [33]. However, [2, Sec. 8.5] cautions that laboratory analysis should confirm monitor readings if such increases are detected.

3.1.4. Normalised Energy Intensity

Background

The *Normalised Energy Intensity* (NEI) method is included in [1, Annex F] and has also been published in [3], [4]. The basis for the approach was introduced in Section 2.4, where different gases indicate different energy-level events. For example, C_2H_2 generally requires more energy to produce than C_2H_6. The NEI method uses their enthalpies of formation from n-octane as a scaling factor for each gas to account for this where applicable [4, Sec. C]. For gases produced by cellulosic materials, the enthalpies of formation from glucose were used instead [3, Sec. D]. The enthalpies of formation values were shown in Table 2-4.

The NEI method is mentioned in [1, Sec. 6.1] where it states "other DGA interpretation procedures exist.... For an example of an alternative approach, see Annex F" immediately after outlining the *DGA Status* levels which it states are to "classify the DGA results, not the transformer condition", indicating it a *Screening* tool. However,

[1, Annex F] is titled "Evaluation of fault severity". It is therefore not clear if it considers the NEI an alternative for *Screening* as well as *Fault Severity*.

The NEI method's publication across numerous sources complicates interpretation due to potentially conflicting information. While [1, Annex F] is the most recent prominent publication, the other sources provide more detailed information. Each source is interpreted separately in chronological order, followed by a consolidated interpretation. The guidance is split into that from [1, Annex F] and those from [3], [4], termed "IEEE Standard Guidance" and "IEEE Journals Guidance", respectively.

Interpretation of IEEE Journal (2012) Guidance [3]

It states that "a practical DGA fault detection and severity ranking method could be based on EWMEA alone or EWMEA with hydrogen and acetylene, if those are available", [3]. *Energy Weighted Methane, Ethane, and Acetylene* (EWMEA) is defined the same as NEI_{3oil} mentioned in [1, Annex F]. The "energy weighted" component is referring to the enthalpies of formations discussed for Table 2-4. One confusing aspect is that EWMEA already includes C_2H_2, but it is assumed that H_2 and C_2H_2 could still be tracked separately in addition to EWMEA.

These are compared against 90th percentiles of the reference population. The DGA scoring method in [3, p. 557] is interpreted as: if the EWMEA is less than the 90th percentile, it is assigned **Ł1**, irrespective of the rate of change in EWMEA. If the EWMEA is greater than 90th percentile but the rate of change is less than the 90th percentile, it is assigned **Ł2.** If both EWMEA and its rate of change are above their 90th percentiles, then it is assigned **Ł3**. However, it is difficult to know which population parameters to use from Table 9 from [3, p. 557] as they are simply labelled "A", "B", and "C" throughout, without any further descriptors. Therefore, the mean values for each are used, and the EWMEA is recalculated as shown in Table 3-3, although this is not ideal. The values for C_2H_2 for population "B" were quite strange with a 90th percentile of 1 ppm whilst the monthly increase was 1.1 ppm. It is unclear how this could be sustainable. It is assumed that the noise from measurements is contributing to this increased value though it does not explain why the other two populations had lower values for the rate of change despite having higher values for the absolute values.

Table 3-3: 90th Percentile Limits for Concentrations and Rates of Change

	Gas Concentration Levels [PPM]						Combined Metric [kJ/kL]
Units	**H_2**	**CH_4**	**C_2H_6**	**C_2H_4**	**C_2H_2**	**CO**	**EWMEA (NEI_{3oil})**
PPM	**87**	89	98	66	**2**	667	**185**
ΔPPM/month	**4.3**	2.5	2.5	1.5	**0.7**	17.3	**7.0**

Source: Values derived from [3, Fig. IX]

Table 3-4: Empirical Limits for Concentrations and NEI

Limits	**Gas Concentration Levels [PPM]**								**IEEE Limit* [PPM]**	**Combined Metric^ [kJ/kL]**	
Perc-entile	**CH_4**		**C_2H_6**		**C_2H_4**		**C_2H_2**		**C_2H_2**	**NEI (NEI_{oil})**	
	Low	**High**	**Low**	**High**	**Low**	**High**	**Low**	**High**	**Combined**	**Low**	**High**
80th	–		–		–		–		–	0.51	0.20
90th	72	18	120	18	120	18	0	–	2	1.02	0.39
95th	120	37	227	37	227	37	1	–	10	1.87	0.72
98th	221	102	433	102	433	102	6	–	36	4.00	1.98

*: IEEE Limit based on 2008 version of IEEE Guidelines [78] and were used for C_2H_2.

^: This was calculated directly using the distribution of NEI values.

Source: Values derived from [4, Figs VI, VII, VIII]

Interpretation of IEEE Journal (2015) Guidance [4]

A notable change in [3, p. 557] is that H_2 is no longer recommended. It states that non-fault processes, such as electrolysis of water, can also generate H_2, and that H_2 DGA measurements have poor reproducibility. As per [4, p. 1943], "to summarize, we recommend inclusion of only the hydrocarbon gases methane, ethane, ethylene, and acetylene for the calculation of NEI..."; these gases match the NEI_{oil} in [1, Annex F].

Table 3-4 combining Tables 6, 7, and 8 from [4, p. 1944] show their limits. Following their values provided, the NEI levels was based on the percentiles of the combined gases and not the constituent gases. By that, it is meant that it was not the 90th percentile for each gas combined as per the NEI formula to give the 90th percentile for the NEI. The 80th percentile included is relevant to the subsequent guidance.

Within their methodology, they explored ostensibly matching the sensitivity of the IEEE Guidelines by defining "new NEI limits with percentiles corresponding in the same way to the values (0.20, 0.10, 0.04) in the ALL row of the HC gas method, that is the 80th, 90th, and 95th percentiles of NEI". The "HC" gas method looked at each gas independently and took the maximum value, mirroring the previous version of the IEEE methodology, [78]. Updating this to the new system used by the IEEE Guidelines is not trivial as in the older approach, the gassing rate was not used to determine the output [78, Sec. 6]. Furthermore, that methodology relied primarily on the *Total Dissolved*

Combustible Gases (TDCG) [78, Sec. 6]. Since [4] considers only gas concentrations, 𝕋1–2 could represent L_1 and L_2, producing a similar scoring system between **Ł1–Ł3**.

In general, this is not being interpreted as recommended best practice, but rather an approach to create a point of comparison used within the paper. For different datasets, the percentiles would likely differ. Additionally, the IEEE Guideline's sensitivity is an arbitrary target that also depends on their definition of sensitivity. Additionally, it is not clear why the value 0.04 rather than 0.05 is used for the 95th percentile.

Interpretation of IEEE Methodology (2019) Guidance [1, Annex F]

It is stated in [1, Annex F] that NEI_{oil} and NEI_{paper} are useful for oil and paper / solid insulation, respectively. These are defined in their equations F.1 and F.2, shown in Equation (2) and (3), respectively.

$$NEI_{oil} = (77.7 \cdot [CH_4] + 93.5 \cdot [C_2H_6] + 104.1 \cdot [C_2H_4] + 278.3 \cdot [C_2H_2]) \,/\, 22400 \quad (2)$$

$$NEI_{paper} = (101.4 \cdot [CO] + 30.2 \cdot [CO_2]) \,/\, 22400 \quad (3)$$

where the gas values in ppm, corrected to standard temperature and pressure (273.15 K and 101.325 kPa), are multiplied by their previously discussed enthalpies of formation, giving an equivalent unit of [kJ/kL]. It states that if significant C_2H_6 stray gassing is suspected, then a variant termed NEI_{3oil} can be considered. This metric simply ignores C_2H_6 in Equation (2). As per [4, Sec. C]:

*"to calculate NEI, the concentration (**µ**L/L)—ppm by volume) of each hydrocarbon gas in multiplied (1 L)/10^6 **µ**L) and then by (1 mol)/(22.4 L) to convert the numerator to moles, then multiplying by (10^3 L/kL) converts the denominator from L to kL. The resulting quantity (mol/kL) is multiplied by the enthalpy of formation (kJ/mol) to obtain kJ/kL for that gas. The sum of the kJ/kL quantities for the four hydrocarbon gases is the NEI."*

It is stated in [1, Annex F] that action may be determined based on outputs of *Fault Type Identification* and the NEI. However, it is not clear how to implement this. [1, Annex F] also states that "separate attention should be paid" to CO, CO_2, and C_2H_2. However, implementation details are not elaborated. As per [1, Annex F], "if NEI_{paper} is increasing, especially if the CO_2/CO ratio is also significantly decreasing, there may be a fault affecting insulating paper", but what constitutes a "significant" decrease in the

CO_2/CO ratio is not defined. Similarly, there is no guidance regarding how to make use of the tracked C_2H_2. The only note regarding limits was, as per [1, Annex F]:

"Experience with this NEI-based method at a large US electric utility suggests that an NEI_{oil} increment of 0.5 or an NEI_{3oil} increment of 0.3 over any time interval should raise concern for the transformer's condition, and larger increments warrant correspondingly more concern".

3.1.5. Lapworth Scoring Algorithm

Background

The *Lapworth Scoring Algorithm* (LSA), published in [5] in 2002, is still used by some industry asset owners. It provides an unbound score more like the NEI outputs than the IEEE / IEC outputs, making it a useful reference point for the NEI. One challenge implementing it compared to the previous three methodologies is that it is confidential with only partial relevant material published in [5]. Although this thesis had access to said algorithm, no further details beyond the already published material is presented.

Motivation

It is stated in [5] that ratio-based methods have issues when gas levels are very low or at zero, where the *measurement Uncertainty* can become very large. It critiques the use of using percentile-based bounds for limits as applied in the IEEE and IEC approaches, noting that when applied across multiple gases, flagging prevalence is inconveniently high. It states that "the main problem with this statistical approach is that gas concentrations which exceed such levels are really only 'abnormal' in the sense of being 'unusual' rather than necessarily 'unhealthy'. Conversely, there is no guarantee that problems will not be experienced below such levels" [5, p. 139].

Methodology

The LSA is based a dataset that was segregated according to: known faulty TXs; normal transmission TXs; normal generator TXs; and heavily loaded TXs. The gas profiles were then investigated, with the known faulty TXs further split into thermal *Faults* and dielectric *Faults*. Based on the findings, a scoring system was proposed which "translates a DGA result into a composite DGA score reflecting the perceived seriousness of the signature" [5, p. 141]. Despite highlighting the shortcomings of ratio-

based methods, it provides little alternative guidance for *Diagnosis*. Additionally, the source dataset is not publicly available.

Assuming the language used in [5] was precise, then it can be interpreted that the scoring algorithm is the mathematical product of two functions, labelled as 'quality' (Q) and 'strength' (S), respectively, with a greater weighting on the first. The first function considers the gas ratios, while the second considers the absolute gas levels. The relevant excerpt, [5, p. 141], describing the functions are provided here for convenience:
"The scoring algorithm used is a product of both 'quality' (dependent on the gas signature and ratios) and 'strength' (depending on absolute levels) functions, but is strongly influenced by the former. …

For the quality function a simple linear expression summing score contributions from the relative amounts of hydrogen and the four hydrocarbons is used. For robustness, ratios are calculated relative to the methane concentration, which is considered to be the most reliable determinant. Relative to methane, score weightings for hydrogen, ethylene, ethane and acetylene concentrations of 150%, 60%, 20%, and 400% respectively were found to give the best fit to the desired outcome.

For the strength function, monotonically increasing but non-linear functions of the methane and carbon monoxide concentrations are used, the latter being included to take some account of perceived indications of overheated cellulose."

This is interpreted as Equation (4) and Equation (5) for the Q function and S function, respectively.

$$Q = \frac{1.5 \times H_2}{CH_4} + \frac{1.0 \times CH_4}{CH_4} + \frac{0.2 \times C_2H_6}{CH_4} + \frac{0.6 \times C_2H_4}{CH_4} + \frac{4.0 \times C_2H_2}{CH_4} \quad (4)$$

$$S = h(f(CH_4), g(CO)), \quad (5)$$

where f and g are "monotonically increasing but non-linear functions" for CH_4 and CO, respectively. Each gas value in Equation (4) and Equation (5) would be expressed in ppm. h is the unspecified function relating the two, giving Equation (6):

$$LSA = j(Q, k) \times S, \quad (6)$$

where k is an unknown weight applied via the unknown function, j, to account for the "strongly influenced by the former" statement in the description of the LSA. The main guidance regarding interpretation in [5] indicated three incremental limits: 30, 60, and 100, representing *Typical*, *Minor*, and *Major* issues, respectively.

Issues with Interpretation

Looking at [5, Fig. 6] and Table 4 from [5, p. 144] retabulated in Table 3-5, respectively, the *LSA* equation seems challenging to derive. Although the *LSA* outputs are provided, and Q can be calculated via Equation (4), there remains too many unknowns to readily solve. The "monotonically increasing" functions of f and g, as well as their combining function, h, in Equation (5) are all unknown. Furthermore, both the weighting factor, k, and the weighting function, j, in Equation (6) are also unknown. The simplest interpretations will fail to accommodate the changes between 23/07/98 and 31/07/98. At this point, both *LSA* and Q increased, whilst both CH_4 and CO also increased. This would require more complicated interpretations. The provided interpretation above is therefore insufficient to readily reproduce the LSA.

Table 3-5: DGA Results and Derived Functions

	Gas Levels [PPM]							Score^	Quality°	Quotient
Date	**H_2**	**CH_4**	**C_2H_6**	**C_2H_4**	**C_2H_2**	**CO**	**CO_2**	**LSA**	**Q**	**LSA / Q**
16/04/1996	10	14	5	16	0.4	140	757	30.2	2.94	10.27
14/08/1997	11	27	9	28	0.7	271	1463	23.3	2.40	9.71
20/04/1998*	14	19	4	18	0.6	64	576	69.3	2.84	24.40
22/07/1998	40	62	12	64	2.9	112	823	96.8	2.81	34.45
23/07/1998	38	**87**	23	131	5.1	**100**	1393	**108.4**	**2.81**	**38.58**
31/07/1998	109	**125**	25	152	5.7	**180**	1142	**121.0**	**3.26**	**37.12**

*: Resistor fitted into core earth circuit and main tank de-gassed before 20/04/1998 result.
^: Score estimated from Equation (6).
°: Quality function estimated from Equation (4).

Source: Gas values from [5, Table 4]

Comparison to Normalised Energy Intensity

Looking specifically at the example shown in Table 3-5, the outputs of the LSA can be compared to the NEI as shown in Fig. 3-5. In black is the LSA with its suggested limits shaded in yellow, orange, and red, representing escalating severity, respectively. The blue and grey show the NEI_{oil} and NEI_{paper}, respectively. The suggested limit by [1, Annex F] for NEI_{oil} is shaded in blue. The results show a consensus near the tail end of the sampling with an escalating severity for both metrics exceeding their limits. However, a fundamental difference between the NEI and LSA approach is highlighted by the third sample in April 1998. At this point, the TX was degassed which reduced the absolute gas levels and thus reduced the NEI metrics whereas the LSA metric substantially increased. This is because LSA uses CH_4 to normalise outputs to account for gradual gas accumulation. In practice, a reviewing engineer would consider the context of the circumstances causing the spike in the LSA.

Another difference is that LSA does not use CO_2 within its metric. This is defensible when considering sources such as Appendix C.5 of [62] which seem to show failure rates are not affected directly by CO_2 levels, however it does point to a potential lack of sensitivity to paper involvement. For example, the NEI_{paper} is rising between the first two samples as the carbon oxide levels double, but the LSA does not increase. Therefore, although LSA and NEI are being compared for potentially fulfilling the same role of *Screening* and/or *Fault Severity*, they should not be considered equivalent regarding how the accomplish that goal.

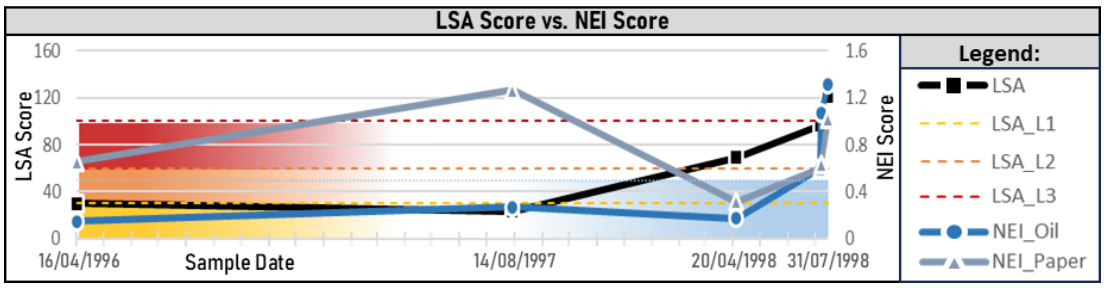


Fig. 3-5: Indicative Comparison of Lapworth Scoring Algorithm to NEI Score

3.2. DGA Uncertainty Application Methodologies

3.2.1. Role of Measurement Uncertainty

Measurement Uncertainty is a crucial component of overall *Uncertainty,* but its scope in relation to CMA can be difficult to precisely define. This is largely due to there being varying approaches to express and quantify *Measurement Uncertainty*. Even if sharing a conceptual approach, as per [79, Sec. 4], "the evaluation of measurement uncertainty is neither a routine task nor a purely mathematical one". Arguably, the *Guide to the Expression of Uncertainty in Measurement* (GUM) and its companion document, the *International Vocabulary of Metrology* (VIM), are the most authoritative sources on this topic. In 1981, the *International Committee for Weights and Measures* (CIPM) tasked ISO with developing a detailed guide to harmonise the expression of *Measurement Uncertainty* [30, p. vi].

Currently, the *Joint Committee for Guides in Metrology* (JCGM) is responsible for maintaining GUM and VIM [79, p. iv]. GUM states that only JCGM publications are authoritative and cites from BIPM when referring to itself [79, p. ii]. This thesis assumes that the versions of GUM and VIM adopted as ISO/IEC Guide 98 [30], [32], [45], [79], [80], [81] and ISO/IEC Guide 99 [82], respectively, are equivalent given that ISO and

IEC are members of JCGM. Unfortunately, the naming schema have historically been unintuitive. Annex ZZ in [79, pp. 12–13] maps both the original and newer JCGM naming schema for GUM and VIM to the current and planned ISO/IEC Guide 98 / 99 versions. This thesis uses the planned naming schema, for example, "ISO/IEC Guide 98-3:2008/Suppl.1:2008" (or "JCGM 101") will be referred to as "Part 7" of GUM, referring to "ISO/IEC Guide 98-7" (or "JCGM GUM-7").

Annex ZZ in [79, pp. 12–13] shows that GUM is expected to consist of 12 Parts, but only 6 have been published since the project began. This is indicative of the potential breadth and depth of this topic. GUM is intended to establish "general rules for evaluating and expressing uncertainty in measurement" [79, Sec. 1]. The concept is to develop a functional relationship, or *Measurement Model*, between relevant inputs and the obtained output, i.e., the measured quantity value(s) attributed to the measurand of the measurement. The *Measurement Uncertainty* expresses the interval of values that could satisfy the assumed conditions of the *Measurement Model*. Note that this phrasing intentionally steers clear of the concept of estimated deviation from the 'true value' due to 'random' and/or 'systematic' errors. Although Appendices D and E in [30] and [82, Sec. 0.1] justify the rationale in detail. As per Appendix E.5.3 in [30], "in practice, the difference in point of view does not lead to a difference… [in output]". The *Measurement Uncertainty* is typically expressed by providing a *Coverage Interval* expected to represent the range of values covering a given *Coverage Probability*. The most common alternative would be to provide a *Standard Uncertainty* associated with the expected value [79, Sec. 4.3].

GUM classifies inputs as *Type A* if statistically derived from observations and *Type B* otherwise [30, Sec. 0.7]. In practice, *Measurement Uncertainty* calculations do not inherently differentiate between them [30, Sec. 0.7] and, as per Appendix E.3.7 in [30], GUM considers them simply "convenient labels" that can sometimes help communicate ideas. The *Measurement Model* should also incorporate relevant correlations between inputs where significant [30, Sec. 5.2], [79, Sec. 4.7]. In practice, quantifying these correlations can be difficult, however, as per [30, Sec. 5.2] "fortunately, in many cases, the effects of such influences have negligible interdependence and the affected input quantities can be assumed uncorrelated". Part 6 of the GUM explores the intricacies in

developing a *Measurement Model* further, and Part 8 of the GUM explores cases where there is more than one output quantity associated with said *Measurement Model*.

Assuming a suitable *Measurement Model*, *Measurement Uncertainties* are evaluated through the *Propagation of Distributions* [45, Sec. 5.2]. A simplified analytical approach is described in Part 3 called the *GUM Uncertainty Framework* (GUF) based on what it calls the *Law of Propagation of Uncertainty* (LPU) [32, Sec. 5.4]. However, GUF relies on certain assumptions to be valid, otherwise, Part 7 discusses the alternative analytical propagation of distributions with the general assumption that it will eventually be evaluated via MCM. A counter-intuitive problem with the analytical simplification provided by GUF is that it can be hard to validate its applicability, and the primary recommended validation approach is via MCM. A more detailed comparison is provided in [45, Sec. 5]. Guidance regarding what information should be included as the final output is provided in [30, Sec. 7], generally summarised as the information necessary to justify and repeat the process of calculating the output. The *Uncertainty Budget* is defined in [82, Sec. 2.33] as a "statement of a measurement uncertainty, of the components of that measurement uncertainty, and of their calculation and combination". Sometimes, a *Sensitivity Coefficient* is also included that attempts to isolate the rate of influence an input has on the output.

Given the context and scope of *Measurement Uncertainty* as described in GUM, there is therefore a clear link between it and *Uncertainty* in CMA. One might consider constructing a *Multi-Stage Measurement Model* that incorporates both the measured DGA value, and the measured percentile-based limits used in, for example, the IEEE *Screening* methodology [1, Sec. 6]. A more challenging example is the *Alarm* limits in the IEC *Screening* methodology [2, Sec. 8], which it states could be tied to PoF. However, much like the construction of a TAI discussed in Section 2.2, the (lack of) availability of information required to quantify these relationships often renders this approach impractical. More typically, *Measurement Uncertainty* is considered within the role of *Conformity Assessment*. *Conformity Assessment* is broadly defined as checking whether a value *Conforms* with *Specified Requirements* [29, Sec. 4.1]. Part 4 of the GUM [80] focusses on a particular facet of *Conformity Assessment* called *Inspection*, where it states that the "determination that a product fulfils a specified requirement relies on a measurement as a principal source of information" [80, p. vii].

ISO 10576 also looks at a similar scope but characterises the differences as it examining the "conformity assessment from a frequentist perspective. ISO/IEC 98-4 examines conformity assessment from a Bayesian perspective" [83, p. v]. ISO 10576 also references PD ISO/TR 13587 [84] which specifically compares the two perspectives as well as a third called the *Fiducial Approach*. It is beyond the scope of this thesis to compare the different approaches, but it should be noted that [84, Sec. 13] concluded that both the interpretation and the numerical results obtained differed between the approaches.

One aspect that [80] and [83] seem to agree on is in generally regarding the assigned limit as absolute. It is stated in [80, p. viii] that limits are based on "business or policy decisions" and are not necessarily metrologically traceable. Factors influencing this decision-making process was covered in Chapter 2 of this thesis. It is stated in [83, Sec. 4] that "measurement uncertainty should neither explicitly nor implicitly be referred to in the designation of the limiting values", but instead "when comparing a measurement result with the limiting values, it is necessary to take into consideration the measurement uncertainty of the result" [83, Sec. 5]. When discussing *Uncertainty* specifically for DGA—including the use of the methodologies in [1], [2]—[66, Sec. C] also advised that limits "should be treated as precise numbers in all calculations and comparisons involved in decision procedures that employ them". Therefore, this thesis assumes *Measurement Uncertainty* to extend into neither the selection of the limit nor the interpretation of the consequence for exceeding the limit.

Assuming *Measurement Uncertainty* is applied solely to the measurement, there are still various ways to assess *Conformance*. Using an *Acceptance Interval* is discussed in [80, Sec. 1] which adjusts the *Tolerance Interval* by accounting for *Measurement Uncertainty*. This is done by attempting to balance the risk and associated consequence of misidentified *Conformance* and *Non-Conformance*. In their approach, a quantity is either accepted or rejected. A *Two-Stage Procedure* is defined and recommended in [83, Sec. 6]. If an obtained measurement with its associated *Measurement Uncertainty* straddles the *Tolerance Interval*, then it suggests a repeat measurement to be done. Then, as per [83, Sec. 6.2], "determine an appropriate combination of the two measurement results to form the final measurement result together with the uncertainty of that result". If the results still straddle the *Tolerance Interval*, then the

conclusion would be an *Inconclusive Test* [83, Sec. 7]. Arguably the simplest approach is outlined by PD IEC Guide 115 [85], which is intended to provide practical guidance to electrical safety testing on the application of ISO/IEC 17025 [86] which specifies the requirement for competence of "testing and calibration laboratories". The relevance of [86] being it is referenced in [68, Sec. 10] as "Guidelines for drafting the [DGA] report in terms of quality assurance". It is argued in [85, Sec. 4] that test methods have "maximum permissible measurement uncertainty expected to be achieved when the method is used", and that developed safety standards will have accounted for this in their limit settings. It then states that "conformance decision is made without applying the measurement uncertainty", calling this *Simple Acceptance* [85, Sec. 4.3.3].

It is therefore argued that no consensus currently exists on the 'correct' application of *Measurement Uncertainty* in the context of DGA of TXs for CMA. The following Sub-Section will review existing Standards-Based literature relating to IEEE C57.104 [1] and IEC 60599 [2] to determine whether they contain a specific methodology for *Measurement Uncertainty* applicable to this thesis.

3.2.2. Standards-Based Literature Overview

Scope

As per [32, Sec. 6], "regulations, legislation or contracts can contain stipulations concerning the measurand, and often these documents specify a measurement to the relevant extent, for instance, by reference to an international standard (such as ISO or IEC) or OIML recommendation". Both IEEE C57.104 [1] and IEC 60599 [2] provide guidance on DGA interpretation, but they lack detailed information on utilising *Measurement Uncertainty*. Therefore, there is a motivation to find guidance on how to implement this aspect. This Section identifies relevant documents and justifies their interpretations for developing a Standards-driven methodology. Fig. 3-6 attempts to map these documents against their indicative scopes and makes apparent the non-trivial nature of the task.

Since both IEEE C57.104 [1] and IEC 60599 [2] list IEC 60567 [68] as normative, they share largely the same basis. Nevertheless, [1] references US Standards (ASTM) more than [2], and the potential consequences of this should be investigated in further work. IEC 60567 provides guidance on analysing/quantifying gases, and it considers

normative: IEC 60475 [67] for sampling oil, and the ISO 5725 series [35], [36], [37], [38], [39], [40] for *Accuracy* of measurement methods and results. [68] also mentions ISO/IEC 17025 [86] for reporting *Quality Assurance* but does not list it as normative. As per Appendix E of [69], ISO/IEC 17025 is relevant for laboratory accreditation, and it mentions both the ISO 5725 and the GUM series (ISO/IEC Guide 98) for conveying *Uncertainty*. ISO/IEC 17025 and GUM rely on VIM [82] for normative vocabulary whereas the ISO 5725 series relies on the ISO 3534 series [87], [88]. However, these cross-reference VIM where applicable and so, are assumed equivalent.

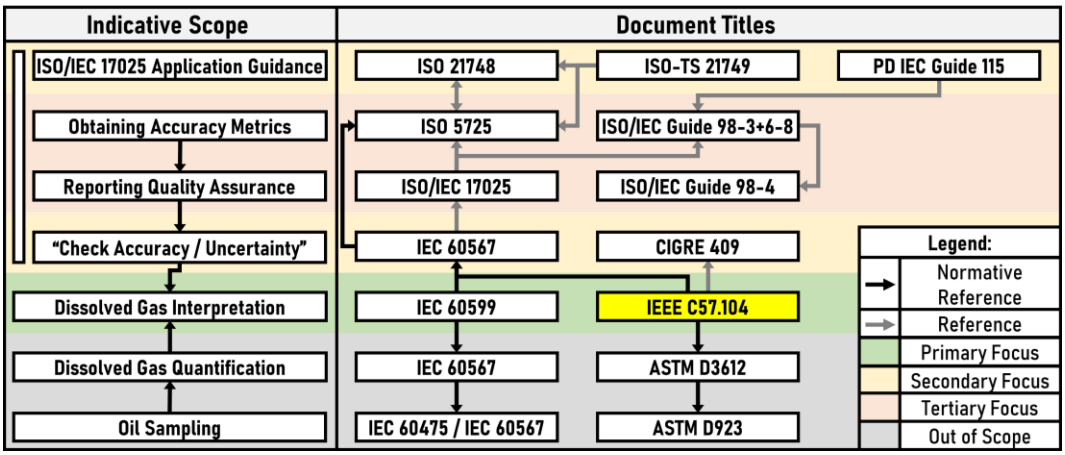


Fig. 3-6: Reviewed documents assumed relevant to Measurement Uncertainty for TX DGA

The advice in [1] and [2] does not provide detailed information on *Measurement Uncertainty*. Both [1] and [2] generally defer to IEC 60567 [68] for obtaining the *Measurement Uncertainty*, which reports *Quality Assurance* via ISO/IEC 17025 [86]. PD IEC Guide 115 [85] is a potential candidate guide on implementing this aspect, as it considers the application of GUM specifically for [86] within the scope of "electrical safety testing conducted within the electrotechnical sector" [85, Sec. 1]. However, IEC 60567 defers to the ISO 5725 series for the methodology to obtain the *Measurement Uncertainty* values. This makes ISO 21748 [89] and ISO/TS 21749 [90], advising on the use of the ISO 5725 series, two further candidate documents. Especially given that [89] explicitly references [86]. In a more general context, GUM Part 4 itself considers the application of GUM for *Conformity Analysis*, and ISO 5725 Part 6 [39] advises on the "use in practice of accuracy values" obtained via the series. For reference, the ISO 5725 is in 6 Parts: Parts 1–4 cover experimental designs and conceptual basics, and Part 5 provides details on some alternative methods.

Fig. 3-6 attempts to map these documents against their indicative scopes. The volume of potentially relevant content to consider can itself present a barrier to implementing a Standards-driven methodology. Shading is used in Fig. 3-6 to indicatively represent the relative relevance to this thesis. The primary focus of this thesis is DGA interpretation, as covered in [1] and [2], with a secondary focus is on *Uncertainty*, for which both methodologies deferred to IEC 60567 [68]. However, to aid with using the information in [68], which is expected to be obtained via the ISO 5725 Series and reported via the ISO/IEC 17025 [86], *Application Guidance* documents are also considered as a tertiary focus. Although ISO/IEC 17025 [86], GUM, and ISO 5725 series are important, they are considered too detailed and broad in their scope. Perceived relevancy was also influenced by the assumption that the decision-makers are mostly using DGA information obtained by third parties. There are therefore many aspects not under the direct control of the assessor that are thus deemed less relevant to this thesis.

Expected Information

For laboratory-based DGA, IEC 60567 [68] gives insight as to what information might be provided for DGA interpretation. IEC 60567 provides instructions for reporting results and includes two relevant notes: "when available, it may be useful for diagnosis purposes to indicate the average accuracies obtained by the laboratory at these gas levels with the analysis procedure used (see 9.3.4)" and "guidelines for drafting the report in terms of quality assurance can be found in ISO/IEC 17025" [68, Sec. 10].

ISO/IEC 17025 [86, Sec. 5] lists several potential metrics, including *Accuracy*, *Repeatability*, *Reproducibility*, *Limit of Detection* (LoD), selectivity, linearity, robustness, and cross-sensitivity. The first four are mentioned in [68, Sec. 9] with the addition of what it calls *Intra-Laboratory Reproducibility*. The wording in [68] suggests that only LoD [68, Sec. 9.2] and *Repeatability* [68, Sec. 9.3] have associated requirements. The requirements for LoD are given in Table 5 of [68, Sec. 9.2] for the gases. The requirements for *Repeatability* are given in [68, Sec. 9.3], but this thesis assumes that *Repeatability* conditions would not be met during typical DGA *Screening* and thus are not used—this will be justified in this Section. For *Accuracy*, [68, Sec. 9.3] states "examples of accuracies that can be obtained using the overall experimental procedure are given in Table 6". Often, when [68] is cited, it is in the context to refencing a ±15% *Accuracy* at medium concentration levels, but the only explicit

recommendation given is that each laboratory determines its own *Accuracy* [68, Sec. 9.3]. For *Reproducibility*, [68, Sec. 9.3] characterises it into *Inter-Laboratory Reproducibility* and *Intra-Laboratory Reproducibility*. [68, Sec. 9.3] states that "inter-laboratory reproducibility has been evaluated by CIGRE as around ±20 % at medium concentration levels" but provides no values for *Intra-Laboratory Reproducibility*.

It is stated in [68, Sec. 9] that *Repeatability*, *Reproducibility*, and *Accuracy* are defined in the ISO 5725 series. Part 1 of ISO 5725 [40, Sec. 3.5] cites [88] for its definition of *Accuracy*, stating it is the "closeness of agreement between a test result … and the true value". An important note is the *Accuracy* includes *Precision*, and thus the stated value will differ depending on whether the test result is based on one or more measurements. If the 'true' value of the parameter, μ, is estimated by $\hat{\mu}$, [90, Sec. 5] states that the bias, δ, would be given by Equation (7), which it states is typically termed a *Correction* when an estimate of it is available. If related to *Accuracy*, bias can also be called *Trueness*.

$$\delta = E[\hat{\mu}] - \mu, \tag{7}$$

For a single measurement, where $E[\hat{\mu}] = \hat{\mu}$, being compared to a known value, the bias can be considered the error. If attempting to estimate the bias in the estimation of $\hat{\mu}$, [90, Sec. 5] states the average error can be used as per Equation (8). [90, Sec. 5] states that if these errors are assumed random, then a probability distribution such as a $\mathcal{N}$ distribution can be used. [90, Sec. 5] says that a "zero" *Correction* is often assumed if the errors are clustered about zero. [90, Sec. 5] states that if the *Corrections* are normally distributed or many *Corrections* are available, then the standard deviation of the sample mean can be used, which is assumed as per Equation (9).

$$\hat{\delta} = \frac{\sum_{i=1}^{n} \hat{\delta}_i}{n}, \tag{8}$$

$$s_{\hat{\delta}}{}^2 = \frac{\sum_{i=1}^{n} \hat{\delta}_i{}^2}{n-1}. \tag{9}$$

If there is not enough information, [90, Sec. 5] says a zero-bias uniform distribution can be used, where the interval is bound by $\hat{a}$ given in Equation (10), and standard deviation given by Equation (11). Equation (11) was modified to remove a $\sqrt{n}$ factor assumed related to the standard error.

$$\hat{a} = \frac{n+1}{n-1}\left(\frac{max\{\hat{\delta}_i\} - min\{\hat{\delta}_i\}}{2}\right), \tag{10}$$

$$s_{\hat{b}} = \frac{n+1}{(n-1)\sqrt{3}}\left(\frac{max\{\hat{\delta}_i\} - min\{\hat{\delta}_i\}}{2}\right), \quad (11)$$

[90, Sec. 5] states that if a non-zero value for the bias is suspected, then measurements should generally be corrected using this value as the *Correction* such that the remaining error is assumed to again be zero-bias. This means that when considering *Accuracy*, it often becomes equivalent to using only measures of *Precision*. For example, [39, Sec. 4] states that "in the absence of specific knowledge of the laboratory component of bias", the equivalent of Equation (12) can be used for comparing a laboratory measurement to a known *Reference Material* (RM):

$$\sigma_{\bar{y}-\mu}{}^2 = \sigma_L{}^2 + \frac{\sigma_r{}^2}{n}, \quad (12)$$

where $\bar{y}$ is the average value measured under *Repeatability*. σ_r and σ_L are the *Repeatability* and *Inter-Laboratory* standard deviations, respectively. These are actually estimates but based on a full *Precision Experiment* as set out in the ISO 5725 series. [39, Sec. 4] argues that these values are as near the true value that will be practically obtained and for clarity, the typical s term will be kept for when only a few samples are used to estimate the term. If a single sample was taken, Equation (13) is the equivalent to using *Reproducibility*, σ_R. This is the same as an estimate for the *Uncertainty* in the result as given in [90, Sec. 5], shown in Equation (13). Again, if a bias is assumed, then each value of y would be expected to first be *Corrected*. There is an added nuance to this: Equation (14), from [89, Sec. 5.3], shows the general statistical model assumed in ISO 5725 series.

$$u^2(\bar{y}) = u^2(\delta) + \frac{s_y{}^2}{n}, \quad (13)$$

$$y = \mu + \delta + B + \Sigma\, c_i x'_i + e, \quad (14)$$

where y is the measurement result, μ is the unknowable true value, δ is the "bias intrinsic to the measurement method", B is the laboratory component of bias, $\Sigma\, c_i x'_i$ are the summated "effects subject to deviation" that were not captured within the collaborative study, and lastly, e is the random error term under *Repeatability* conditions. If the *Trueness* of the measurement model, δ, is thought to be known, then [89, Sec. 7] states it can be used for *Correction*. However, [36, Sec. 7] explicitly states that although laboratories should made aware of their laboratory component of bias, B, it should not be used for *Calibration* purposes or any *Corrections*. Given that [68,

Sec. 9.3] gives no indication to the value of δ, this thesis assumes a "zero" *Correction*, and that Equation (15) is the applicable scenario for *Accuracy* assuming a single measurement is taken.

$$u(y) = \sigma_R, \tag{15}$$

[39, Sec. 4] states that when comparing measurement(s), the ISO 5725 series typically assumes a probability level of 95%. As such, the thesis assumes that the value given in [68, Sec. 9.3] represented a $\mathcal{N}$ distribution with a *Coverage Factor*, k, of 1.96. ISO 5725 uses the term *Critical Distance*, f, [39, Sec. 4]. This gives Equation (16) with then the *Inter-Laboratory Reproducibility* interpreted as per Equation (17):

$$f\sigma_R = \pm 15\%, \tag{16}$$

$$\sqrt{2}f\sigma_R = \pm 20\%. \tag{17}$$

For reference, $15 \times \sqrt{2} \approx 21.2\%$ and $20/\sqrt{2} \approx 14.1\%$. This was assumed within tolerance indicated by use of the word "around" in [68, Sec. 9.3] given the phrasing in [66, p. 23] which will be explored in greater detail later. However, it could also be interpreted as there being a *Correction* for the shared bias, δ, whose *Uncertainty* affects *Accuracy* but not *Reproducibility*, as shown in Equations (18) and (19).

$$u(y)^2 = u(\mu - \delta)^2 + \sigma_R^2, \tag{18}$$

$$(15)^2 \approx (5.1)^2 + (14.1)^2. \tag{19}$$

Critique of Interpretation

The interpretation explained above used in Chapter 5 of this thesis presents several issues. The information being used is from [68] and as such, it is a *Type B* source of *Uncertainty*, regardless of whether it was itself a *Type A Uncertainty* when first estimated at the source. Appendix A in [85] states that for *Type B* sources, a $\mathcal{N}$ distribution is only applicable if a *Coverage Factor*, k, is given. In the publication related to Section 5.2, [47], a triangular distribution was because the *Coverage Factor* was unknown, but it was assumed that subsequent measurements would be closer to the obtained measurement than to the interval edges. As per Appendix A in [85], "a triangular distribution should be assigned where the contribution has a distribution with defined limits and where the majority of the values between the limits lie around the central point". Section 5.2 then took the stance that as the ISO 5725 series was claimed normative in [68] and explicitly referenced both when defining *Accuracy* and

as the methodology that ought to be employed when obtaining *Accuracy*, it was justifiable to assume its *Coverage Factor* was applicable.

It could also be argued, as per Appendix A in [85], that "a rectangular distribution should be assigned where a manufacturer's specification limits are used as the uncertainty" when a *Coverage Factor* is not given. In the examples provided by Appendix A in [85], for every *Type B* source of *Uncertainty*, including for *Repeatability*, it used a uniform distribution. It is a limitation of this thesis that a uniform distribution was not also covered in Section 5.2. However, in GUF [30], *Uncertainties* are simplified to be expressed as equivalent Student's t-distributions, often simplifying further to $\mathcal{N}$ distributions. An example of this is done in Section 5.2 where an equivalent $\mathcal{N}$ distribution for a triangular distribution is used. If a uniform distribution is preferred, an equivalent $\mathcal{N}$ distribution could be calculated, and its value used in place of what was there in the *Accuracy* as an approximation.

Arguably, the more serious issue with the approach used in Section 5.2 is regarding how the *Uncertainty* of multiple samples was calculated. If only *Accuracy* is available, then there is no alternative to the approach used. However, it is important to recognise that a key requirement of an effective *Measurement Model*, as described in GUM, is that influencing factors should not be shared among variables unless explicitly accounted for through what it calls *Sensitivity Coefficients* and covariances. For example, when considering multiple samples for a delta, or change gas levels, then they may be expected to share influencing factors, such as the laboratory quantifying the dissolved gases. When considering a delta, or a linear regression, the *Accuracy Uncertainty* should have been modified by some *Sensitivity Coefficient* and relevant covariance. GUM [32, Sec. 10] discusses this topic in greater detail. Annex A of [89] also covers the topic with a focus on comparing with the ISO 5725 series, and provides Equation (20) where the *Sensitivity Coefficient*, c_i, is given in Equation (21).

$$u(y)^2 = \sum_{i=1}^{N} c_i^2 u(x_i)^2 + \sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} c_i c_j u(x_i, x_j), \tag{20}$$

$$c_i = \frac{\partial y}{\partial x_i}, \tag{21}$$

where x_i is an input into the *Measurement Model* propagated to estimate the output *Measurement Uncertainty*.

In the ISO 5725 'mindset', this would instead be captured by adjusting the measures of variability to represent the applicable scenario. For example, under the conditions of *Repeatability*, the standard deviation of *Repeatability* would be used. Under the conditions of *Reproducibility*, the standard deviation of *Reproducibility* would be used. It is highlighted in [90, Sec. 5] that "modern instrumentation is exceedingly precise in the short term, but changes over time, often caused by environmental effects, can be the dominant source of uncertainty in the measurement process". It goes on to provide various potential experimental setups to estimate these factors, where they are evaluated via ANOVA. It also suggests either a two- or three-level characterisation:

- Short-term fluctuations (*Repeatability* or instrument *Precision*),
- Intermediate fluctuations (day-to-day or operator-to-operator or equipment-to-equipment, known as *Intermediate Precision*), and
- Long-term fluctuation (run-to-run or stability or *Intermediate Precision*).

If Equations (18) and (19) are the correct interpretation, then Section 5.2 ought to use these values as applicable in the calculations. Otherwise, one could argue this is a shortcoming of the information being provided in [68] by it not giving a value for *Intra-Laboratory Reproducibility* when it would be very relevant for DGA *Screening*, where the changes in gas levels are of great interest. The 2005 publication [66] discussed the usage of some of these metrics. The scope of [66] is only on *Uncertainty* associated with quantifying the dissolved gas concentrations, and *Blunders* are also excluded as it states they cannot be reliably predicted or evaluated. As was explained for Equation (14), other *Uncertainties*, such as those associated with sampling, could presumably be added. If a single measurement is being evaluated, or multiple measurements taken by different laboratories, [66, p. 23] says *Accuracy* should be used. If multiple measurements are taken by the same laboratory, then it says either *Repeatability* or *Intra-Laboratory Reproducibility* can be used, depending on whether the measurements were taken "over the same day or a short period of time" or over a longer period, respectively. [66, p. 23] does not use *Inter-Laboratory Reproducibility*: this gives credence to the assumption made for using Equation (15) over Equations (18) and (19).

If it is assumed that *Repeatability* was measured in accordance with [68] rather than ISO 5725, and that the measured values were equally representative of all gas values in its range, then this advice seems applicable. It is potentially less applicable with ISO

5725 that has in its definition for *Repeatability Conditions* that "independent test results are obtained with the same method on identical test/measurement items…" [40, Sec. 3]. It is generally assumed that the derived estimated *Accuracy* and *Precision* based on the inter-laboratory tests using ISO 5725 may be applicable to the range of levels it states to support. However, when the discussion is regarding two different gas samples, potentially significantly different enough to give meaning to a measure of a gassing rate, then it requires assuming the biases will cancel out between these two samples. This is fundamentally different from assuming that biases will cancel out between two samples expected to be generally the same. For example, when OLDGA was assessed in [33, Sec. 10], it evaluated *Reproducibility* and not *Repeatability* by "looking at the small fluctuations and width of the baseline when the monitor is installed on a transformer which does not gas, or which gases only very little". The difference in reasoning may be because [68] is discussing specifically comparisons between samples. In the context of [1], where gassing rates are calculated based on samples over 6–24 months, it would require *Intra-Laboratory Reproducibility* to have been estimated in a *Precision* test that adequately represented all the variability to be expected. It would be very easy to underestimate this metric or have it no longer representative of a laboratory's current performance given the timeframes involved.

Establishing relevance of method performance data to measurement results, and continued verification of performance, is discussed further in [89, Sec. 7]. It should be noted even when using a *Gas-in-Oil Standard* (GIOS) sample to estimate performance, the GIOS sample itself may be a source of *Uncertainty*. Annex E of [68] assumes an *Uncertainty* of ±2% related to the GIOS sample. Appendix E of [69] stated that "repeated tests by one manufacturer of commercial GIOS samples have shown that the loss of hydrogen one week after they have been prepared is less than 1 % (<2.5% per month)". They then recommend that if attempting a longer-term study, that "long life" GIOS samples be used.

The inclusion of *Sensitivity Coefficients* would introduce many challenges to the proposed methodology. The fundamental issue is that *Accuracy* and *Reproducibility* are estimates that area both required concurrently when applied to the methodology outlined in [1]. Attempting to assert these estimates in the form of a correlation will have scenarios arise where there are no valid combinations of *Sensitivity Coefficients*

that can satisfy the assumptions. A common example is when a value of 0 ppm is recorded, allowing no scope for correlations to achieve coherent assumptions.

Other Factors Influencing Uncertainty

Sampling Uncertainty

Both the *IEC Specification* [68] and the discussed publications, [66], [70], [71], focussed on the *Uncertainty* associated with laboratories quantifying dissolved gases in a sample. If a portable monitor is used instead, then the former values should be used with caution as the sample is not quantified within a laboratory. The same applies for OLDGA, with the addition that it also does not involve manual sampling: which may be a significant factor to consider for both portable monitors and laboratory analysis. It is often recommended to still send confirmation samples to a laboratory [2, Sec. 8].

Some results from 4 *Round Robin Tests* (RRT) were published in [71]. Its RRT-1 had *Uncertainty* from sampling included whereas RRT-4 instead used a GIOS sample. There are results in [33, Sec. 9] for *Repeatability* of its RRT directly, which seemed to include sampling from a TX. It found that at routine gas levels, which they defined as > 5 ppm, the laboratories averaged a *Repeatability* between ±2% and ±9% depending on the gas. Oxygen was an outlier with ±21% which was stated as commonly contaminated within samples by the atmosphere. However, the overall sample sizes of these examples are too small to differentiate the impact of the sampling *Repeatability* from the quantifying *Repeatability*. For example, Table 1 in [66] estimates average laboratory *Repeatability* at ±7%, ranging between ±1% and ±15% depending on the laboratory.

Additionally, sampling *Uncertainty* involves factors often classified as *Blunders*, such as gas contamination, which are not considered as part of *Measurement Uncertainty*. Arguably, the main source of reliable *Uncertainty* would be from inhomogeneity of samples. The topic of inhomogeneity is mentioned in [89, Sec. 5]. It distinguishes inhomogeneity within a sample from inhomogeneity within the population. It further distinguishes *Uncertainty* in estimating a mean of samples taken from the population as compared to the *Uncertainty* in said estimate being used to represent the population. One challenge in quantifying this is that a TX is a dynamic system and the inhomogeneity between samples may change depending on the time elapsed. Taking two samples in quick succession would provide their relative difference, but not how

they compare to the 'population'. Waiting a longer duration would then bring into question how much the 'true' value changed during the interim.

As per [69, Sec. 4], "it has been mentioned that 90% of DGA values obtained by oil sampling were taken from the bottom oil" where equilibrium is reached quite slowly compared to the top oil. The specifics would depend on the TX design, but [69, Sec. 4] gives some general guidance stating that "tests done on a transformer have indicated a 2-hour delay for gas injected on one side of the bottom oil to reach the other side of bottom oil and the top oil", and that equilibrium should be reached in the case of "low-to-moderate gas formation" defined as under the limit rate levels in [62]. This source of inhomogeneity becomes more significant at higher gas formation rates, especially if partially undissolved gas bubbles are formed in the oil.

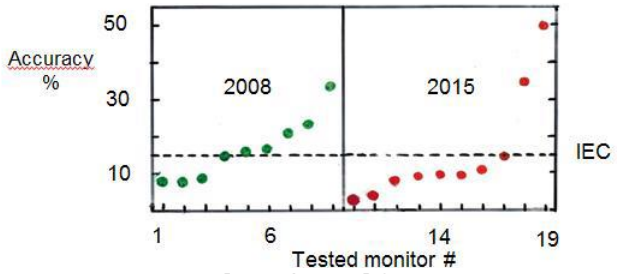
Online DGA Uncertainty

When discussing these methods comparatively, a relationship between *Accuracy* and *Precision* should not be assumed. For example, portable instruments eliminate *Uncertainties* related to transport but introduce other issues, such as less controlled environments and potentially less reliable or *Accurate* instrument [33, Sec. 5]. Their main advantage is the speed at which a result can be obtained which can be paramount for a relatively volatile TX. However, they are generally considered less *Accurate* than laboratory testing [1, Annex G.5]. Another example is that OLDGA monitoring systems similarly remove the need for transportation but also the manual laboratory analysis as well, and though this increased automation can improve the short-term *Precision* of the outputs, other factors may affect its *Accuracy* [33, Sec. 10]. In particular, the challenges associated with maintaining the same degree of calibration as in a laboratory and the potential for sensor drift over time. Another issue is factory calibration on the online monitor may have used a different oil type to that in the specific TX [33, Secs 6, 9]. Some monitors mitigate this via programmable offsets for onsite recalibration to match laboratory results [69, Sec. 6]. Some devices can also automatically recalibrate onsite, but it uses pre-stocked gases that must be periodically restocked [69, Sec. 2]. The result is that online monitors are considered less *Accurate* but more *Reproducible* than a generic laboratory. To further emphasise the distinction between the metrics: Table 7 in [33, Sec. 6] has a list of DGA online monitors with their respective manufacturer's claims regarding *Reproducibility* and *Accuracy* where the model "4810"

has the best *Reproducibility* whilst also having amongst the worst *Accuracy*. This evaluation can change, as per [69, Sec. 6]: "after some time in operation, the accuracy of some monitor models ... have indeed been observed to drift and become poorer".

It is challenging to quantify metrics comparatively generally as they are dependent on the specific laboratory, portable monitor, or online monitor. Furthermore, there are other relevant factors for consideration. For example, the sampling frequency may render laboratory's *Accuracy* metrics irrelevant when considering a fast-occurring *Fault* that develops during the interim. A *Cost-Benefit Analysis* (CBA) is demonstrated in [69, Sec. 2] to justify OLDGA based on [19]'s cited annual failure rates due to cases of typical gas levels followed by fast-occurring *Faults*. As OLDGA technology matures, previously published findings may lose their relevance. For example, [33], [69] both had similar scopes with the latter being the more recent follow-up (2010 versus 2019) and during this short interval, Fig. 3-7 from [69, Fig. 6.1] shows a noticeable improvement in the average *Accuracy* of surveyed online monitors.

Online DGA monitors were tested in [33, Sec. 9], and if taking the weighted average of the models reviewed in their Table 27, the average *Accuracy* for online monitors was estimated at ±18%. *Reproducibility* was estimated in [33, Sec. 10] by "looking at the small fluctuations and width of the baseline when the monitor is installed on a transformer which does not gas, or which gases only very little", stating that: "it has thus been evaluated as ±1-2% for the TM8 and the TNU, ±2-4% for the Tranfix, ±4-10% for the TGM and ±5-10% for the Hydran (< 5% in the more recent units)". As discussed earlier, [66, p. 23] stated that if comparing between samples measured near the same time, then this may qualify as *Repeatability*. The findings in [33] suggest that a good laboratory can achieve higher *Accuracy* than online monitoring, assuming the sampling is done correctly, but that OLDGA is comparable. However, the more recent [69] found the average *Accuracy* of the newer selected monitors is ±13%. No new *Accuracy* values were provided for laboratory testing: it is therefore unclear whether these have made similar improvements in *Accuracy* in this time.



Source: From [69, Fig. 6.1] from CIGRE © 2019

Fig. 3-7: Accuracy of DGA monitors tested by CIGRE

3.2.3. In-depth Review of IEC 60567

Preface and Context

The following work is not reflected in Section 5.2 but may be of relevance to those wishing to use the values given in [68, Sec. 9]. IEC 60567 is sometimes simply called the *IEC Specification* as will be done here. The focus in this thesis will be on the 2011 version [68], however, there are also the 2005 [91] and 1992 [92] versions that may represent what was available at the time of a given publication. It should be noted that there is also a pending 2024 version that may address the points raised here. There will be several criticisms made by the thesis of the newer versions of the *IEC Specification*, [91] and [68], regarding the presentation of the information. Firstly, the critiques are possible only because the information is more accessible: this thesis does not explore the comparative validity of the values in [92] relative to what they purported to represent. Secondly, the intent is not to denigrate either [68] or the values provided within it, but to highlight the current phrasing potentially mischaracterises them. The scope of the critique is purely on the wording used and not on more fundamental discussions of the statistical validity of them which is beyond the capabilities of this thesis. Clearly, given that Section 5.2 relies heavily on the information given in [68], its value is appreciated.

The *IEC Specification* is commonly cited, and as discussed, it is considered normative for both DGA interpretation methodologies: [1], [2]. Some example references to the *IEC Specification* are provided to give context; excluding the last example, these are not cherry-picked but instead represent significant publications. The 2003 publication [71, p. 9] includes in a table headed “average accuracy for different extraction techniques”, values it states are the “accuracies specified for the standard techniques” referencing [92]. Table III of [71] characterises said *Accuracies* as 13% for medium concentration

levels and 35% for low concentration levels. The 2005 publication [66] discusses performance results that seem to be the source of the updated *Accuracy* values of ±15% provided in [91] and [68]. It then states "if the laboratory's own accuracy and reproducibility estimates are not known, the CIGRE averages in (1) can be taken as default values for purposes of data interpretation" [66, p. 23]. The 2010 publication [33, Sec. 9] states "at routine concentration levels, the IEC specification for accuracy is ±15% in order to get a reliable diagnosis" citing [91]. The 2019 publication [69, Sec. 6] mentions "the accuracy requirements of IEC (<15%)" citing [68]. The 2019 publication [1, Sec. 5.2] states that [68] provides "recommendations to have accuracies better than ±15% on DGA results…". The 2022 publication [2, Sec. 6.2] mentions that [68] describes how to calculate DGA *Uncertainty*, also mentioning the same ±15% *Accuracy* and later encourages "using DGA data meeting IEC 60567 specifications for accuracy…" [2, App A.1]. Lastly, significant only to this thesis, the conference publication related to this thesis looking at *Measurement Uncertainty* [47], uses this same ±15% *Accuracy*.

Source Material

Clause 9 of [68], "Quality Control", is the most relevant and is divided into Sub-Clauses:

- verification of the entire analytical system,
- limits of detection and quantification, and
- repeatability, reproducibility and accuracy.

Clause 10: *Report of Results* of [68] also mentions [86] for "Guidelines on drafting the report in terms of quality assurance". Annex E of [68], "Procedure for comparing gas monitor readings to laboratory results", is also relevant and will be discussed.

Sub-Clause 9.1: Verification of the Entire Analytical System

Sub-Clause 9.1 recommends the use of at least two *Gas-In-Oil Standards* (GIOS) samples, one for low concentrations "resembling oils in factory tests" and one for "oils resembling oils from equipment in the field", to "check the quality of the results produced by the analytical system" [68, Sec. 9.1]. It then states the results can be used to compensate for "incomplete extraction and other operational factors". Lastly, it states that "good practice" would have this "procedure at intervals of calibration" at least every six months or sooner if there is a change in apparatus or operating conditions. This is distinct from the inter-laboratory RRTs described by the ISO 5725

as [40, Sec. 7] specifically states that these RRTs should not be used for *Calibration* or *Corrections*. Similarly, [69, p. 42] mentions that laboratories wishing ISO 17025 *Accreditation* require both GIOS samples in proving results *Precise* and *Accurate*, as well as inter-laboratory RRTs based on ISO 5725 at least every year.

Furthermore, this "verification of the entire analytical system" is distinct from *Calibration* of specific equipment which may be more frequent. For example, daily *Calibration* of the chromatograph [68, Sec. 1], or *Calibration* of the headspace extractor that is expected to be done at least once a month if using a GIOS sample [68, Sec. 7]. Some current (2019) practices in DGA laboratories were discussed in [69], stating that *Calibration* of the chromatograph is done "typically every one to four days" [69, p. 37]. It also stated that some labs "successfully recalibrate their DGA laboratory equipment with GIOS sample every 4 months for accreditation purposes, and every month internally, to ensure a very good accuracy of their results.... Others verify their calibration curves with GIOS every day, and use 6 different concentrations of GIOS to prepare their DGA calibration curves" [69, p. 42].

Sub-Clause 9.2: Limits of Detection and Quantification

Table 5 in [68, Sec. 9.2] gives guidance on *Limit of Detections* (LoD) that should be achievable. It differentiates the use-case for *Acceptance Testing* and for *Service Testing*. LoD is defined as the "lowest concentration that can be identified" and *Limit of Quantification* (LoQ) as the "lowest concentration that can be quantified with a reasonable precision and accuracy" [68, Sec. 9.2]. These terms are explained further in [93], stating "if the observed bias and imprecision at the LoD meet the requirements for total error for the analyte ... then: LoQ=LoD". Therefore, it is assumed that the subsequent guidance on *Precision* and *Accuracy* in [68, Sec. 9.3] would be applicable to define LoQ. For reference, when the older version [92, Sec. 9.1] described LoD, it stated "sensitivity is the ability to detect a given gas species with high confidence (e.g. 95%) at very low concentration.... A measure of sensitivity is the detection limit ... it is generally considered that the detection limit for any gas is at approximately twice ... background noise level". This is described as a "traditional and typical approach to estimate LoD... as the mean +2 SD" in [93, p. S50]. Given the values of LoD for hydrocarbons and carbon oxides have remained unchanged since [92, Sec. 9.1], the LoD could potentially be used as a rough approximation for the minimum *Uncertainty*.

The values for "atmospheric gases" changed drastically between versions, from 50 ppm to 2,000 ppm, suggesting a different method for the derivation of their values.

The relationship between LoD and LoQ depends on many factors, and as per [93, p. S52], only LoD being less than or equal to LoQ can be presumed. One issue with how [68, Sec. 9.3] expresses *Repeatability* requirements is that it is based on LoD, S. This could imply laboratories with lower LoDs face more stringent LoQ requirements. For example, when [33, Sec. 8] looks at comparing the performance of OLDGA, they standardise LoD to either the values in Table 5 in [68, Sec. 9.2] or the manufacturer's stated values, whichever is greater, to "get comparable results and to be fair…".

Sub-Clause 9.3: Repeatability, Reproducibility, and Accuracy

This Sub-Clause is the focus of the review in this thesis, and it discusses *Repeatability*, *Reproducibility* and *Accuracy*, where it refers to ISO 5725 for their definitions [68, Sec. 9.3]. Through the incremental changes in IEC 60567 from the 1992 version [92] to the 2005 version [91], this thesis argues that there is a disconnect between the methods and values as would be interpreted via ISO 5725 to those given in this Sub-Clause.

Sub-Clause 9.3.2: Repeatability

Although example values are given for *Accuracy* and *Inter-Laboratory Reproducibility*, [68, Sec. 9.3] reads as having specific performance-related requirements only for *Repeatability* by providing an equation and a "general acceptable value". No requirements are given for *Reproducibility* nor *Accuracy* to which a laboratory is expected to meet.

IEC 60567 defines *Repeatability* as: "related to the differences that are observed when the same oil sample is analysed by the same laboratory over the same day or a short period of time" [68, Sec. 9.3]. The ISO 5725 series uses a stricter definition, explicitly requiring aspects such as the equipment, operator, etc., to remain unchanged [40, Sec. 3]. IEC 60567 uses r for *Repeatability* and R for *Reproducibility*, stating they are "defined in detail in ISO 5725" [68, Sec. 9.3]. This likely refers an older (1986) version of the ISO 5725 series which used r for the *Repeatability Value* and noted that it is shortened to *Repeatability* [94, Sec. 3]. Unfortunately, [68, Sec. 9.3] does not seem to align well with the more current versions of the ISO 5725 series, which currently describes *Repeatability* and *Reproducibility* as types of *Precision* [40, Sec. 3]. It explains

Precision is "usually expressed in terms of imprecision and computed as a standard deviation of the test results" [40, Sec. 3.12] and that "two quantities are required as measured of precision, the repeatability standard deviation … and the reproducibility standard deviation", using σ_r and σ_R, respectively [40, Sec. 5.2]. However, as mentioned in [39, Sec. 4], "normal laboratory practice requires examination of the difference(s) observed between two (or more) test results, and for this purpose some measure akin to a critical difference is required, rather than a standard deviation". As per [88, Sec. 3], the *Repeatability / Reproducibility Critical Difference* (CD) is the "value less than or equal to which the absolute difference between two final values … is expected to be with a specified probability".

These, as per [40, Sec. 3], are defined by the *Repeatability / Reproducibility Limits* which corresponds to "the value less than or equal to which the absolute difference between to test results obtained under [the relevant term] conditions may be expected to be with a probability of 95%". In other words, the *Limits* are the *Critical Difference* when a probability level of 95% is used. Throughout the ISO 5725 series, r represents the *Repeatability Limit* and R the *Reproducibility Limit*, respectively, as seen in Appendix A of [40]. Equation (22) is provided in [39, Sec. 4] for the *Repeatability Limit*, with a derived example usage being to assert Equation (23) must be upheld.

$$r = f \times \sigma_r \times \sqrt{2}, \tag{22}$$

$$r \coloneqq |y_1 - y_2| \leq f \times \sigma_r \times \sqrt{2}, \tag{23}$$

where f is defined as the *Critical Range Factor*, taken as 1.96 for 95 % probability level, and σ_r is the standard deviation of the *Repeatability*. y_i are measurements obtained under *Repeatability* conditions. Clearly, one cannot suppose r and σ_r are interchangeable in this context. The value for σ_r is assumed to be shared amongst the participants of the *Precision Experiment* and so it taken as the average of each laboratory's *Within Laboratory* standard deviations, σ_W.

An equivalent of Equation (24) for medium gas concentrations (>10 ppm) is given in [68, Sec. 9.3], stating that it "means that the repeatability of the laboratory, at 95 % confidence limit, is lower than k times the mean concentration of the gas analysed". It is then stated in [68, Sec. 9.3] that "for low gas concentrations (for example, < 10 µl/l), the required repeatability is given by the following equation: $r = S$ (where S = detection limit, whatever the concentration…". This has been interpreted as Equation (25).

$$r \coloneqq (y_1 - y_2) < k \times (y_1 + y_2)/2, \tag{24}$$

$$r = S, \tag{25}$$

where S is the LoD. The original used A and B in place of y_1 and y_2 to represent measurements: this substitution is repeated henceforth. A "general acceptable value" for k is given as 0.07 if gas concentrations are between 10–1,000 ppm, or 0.10 if above, in [68, Sec. 9.3]. It seems clear that Equation (24) aligns with Equation (23). This thesis speculates Equation (25)'s differing presentation as to Equation (24) is indicative of the misalignment between [68] and [40]. For reference, originally [92, Sec. 9] had Equation (26) in place of Equation (25). Equation (24) did not change.

$$r \coloneqq |y_1 - y_2| < 2 \times S + k \times (y_1 + y_2)/2, \tag{26}$$

where k was taken as 0.15—and as 0.10 if applied to Equation (24) for higher concentration values. Equation (25) could have kept its original structure even with the change from a *Relative* to an *Absolute Uncertainty*. Four functional relationships for *Relative Precision* are outlined in [35, Sec. 8] which it considers likely sufficient to describe many situations. The first two are given in Equations (27) and (28):

$$s_r = b \times m, \tag{27}$$

$$s_r = a + b \times m, \tag{28}$$

where s_r is the estimated standard deviation of the *Repeatability*, and thus here assumed equivalent to σ_r, and m is the "general mean (expectation)". Although note, as per [40, Sec. 5], "the level m is not necessarily equal to the true value μ". If substituting Equations (27) and (28) back into Equation (23), the derivation from Equations (24) and (25) would be Equations (29) and (30), respectively.

$$|y_1 - y_2| \leq \sqrt{2} \times f \times b \times (y_1 + y_2)/2, \tag{29}$$

$$|y_1 - y_2| \leq \sqrt{2} \times f \times (a + b \times (y_1 + y_2)/2), \tag{30}$$

This implies Equations (31) and (32) and where $f \times \sqrt{2} \approx 2.8$ as per [39, Sec. 4].

$$k \approx 2.8 \times b, \tag{31}$$

$$S \approx \sqrt{2} \times a, \tag{32}$$

Equation (31) could be interpreted as combining the minimum *Uncertainty* being represented by LoD, S, which [92, Sec. 9.1] mentioned was traditionally based on background noise, where f ≈ 2. If evaluating Equations (31), b would be equal to 0.025 and 0.036 for when k is equal to 0.07 and 0.10, respectively. This would mean s_r would be equivalent to b as a percentage of m and not k, e.g. ±2.5% rather than ±7%. However,

this thesis instead speculates that the 0.07 and 0.10 given for k in [68, Sec. 9.3] are intended as values for b in Equation (31). One point of corroboration is that using the value in [68, Sec. 9.3] for k accordingly, gives the previous value of k in [92, Sec. 9.2]: $\sqrt{2} \times 0.07 \approx 0.10$.

Annex E was added to the 2011 Version [68], and explained the procedure to evaluate the maximum *Accuracy* of OLDGA by using laboratory results. This was based on [33, Sec. B], using same values and terminology. In it, they mention to "calculate the repeatability (R) of laboratory results as the difference between results for the individual 4 samples and average values (A), and express it as a percentage". However, it is arguably unclear how they calculated their answer in the example. This thesis speculates Equation (33) is indicative of the approach used, based on Equation (10):

$$\hat{a} = \frac{n+1}{n-1}\left(\frac{\underset{i\in n}{max}\{|y_i - \bar{y}|\} - \underset{i\in n}{min}\{|y_i - \bar{y}|\}}{2}\right) \times \frac{100}{\bar{y}}, \tag{33}$$

where $\hat{a}$ is the estimated range of a uniform distribution assumed to represent the *Uncertainty*. Appendix F of [69] revisited a very similar example in 2019 and provided the equivalent of Equation (34):

$$X_r = \frac{\sum_{i=1}^{n}|y_i - \bar{y}|}{n} \times \frac{100}{\bar{y}}, \tag{34}$$

where X is used because it is not clear specifically what aspect of *Repeatability* Equation (34) estimates, although if $\bar{y}$ can be assumed as an estimate of the 'true' value, then X would be analogous to *Mean Absolute Percentage Error* (MAPE) [95]. Equation (35) given in [89, Sec. 7] for estimating a laboratory's mean bias, $\bar{\Delta}_y$, is also similar but does not use the absolute difference. Equation (36) is then the estimated variance of the laboratory bias [89, Sec. 7]. This essentially the same as the equation given in [37, Sec. 6] defining the estimate of the *Within-Laboratory* standard deviation, s_i, which can be considered representative of the general approach to estimate a relevant standard deviation such as the *Repeatability*.

$$\bar{\Delta}_y = \frac{\sum_{i=1}^{n}(y_i - \bar{y})}{n}, \tag{35}$$

$$s_\Delta{}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}. \tag{36}$$

Given the volume of potentially relevant material, mention of Equation (34) and its prescribed usage may well have been missed by this thesis. A brief review of Part 5 of

the ISO 5725 series, dedicated to alternative methods, found only two robust estimation methods called *Algorithm A* and *Algorithm S*, neither of which aligned with Equation (34) [38, Sec. 5]. Most generally, one can say that not using the root squared differences will mathematically make it less sensitive to outliers, a big concern with very low sample numbers, although it could equally be said that with few samples, there is a tendency to underestimate variability if not accounting for degrees of freedom. It may have another statistical implication not investigated here. The more significant implication is that both Appendix B of [33] and Appendix F of [69] applied different equations with significantly differing results to ostensibly represent the same metric, *Repeatability*, and then proceeded to use the derived value in the exact same way in the subsequent calculations. Consider that Equation (33) estimated the interval of a uniform distribution: as per Equation (11), the standard deviation of *Repeatability*, s_r, would be a factor of $\sqrt{3}$ smaller. Furthermore, the lack of mention of a *Coverage Factor* also supports the previous claim that the values for k in [68, Sec. 9.3] are instead intended as values for b in Equation (31). This complicates interpreting values in [68] that are intended to represent *Repeatability*.

This thesis asserts that the dataset discussed in [66, p. 22] from IEC laboratories and from CIGRE laboratories are those used in [68]. It is stated in [66, p. 22] that the average *Repeatability* for the IEC laboratories surveyed was ±7% at medium gas concentrations. This is speculated to be the origin of the change of characterisation from $k = 0.10$ [92, Sec. 9.2] to $k = 0.07$ [91, Sec. 9.3] and is largely driven by the discrepancy in the $\sqrt{2}$ factor used when combining an *Uncertainty* twice via addition in quadrature. In other words, when [92] defined k, it was in reference to difference between two samples, then when [66, p. 22], [91], and [68] defined k, it was in reference to the *Uncertainty* of a single sample. This detail is not clearly conveyed in IEC 60567-2011 [68]. Furthermore, [66, p. 22] states that this is the average *Repeatability* of the laboratories. It is not clear if this was calculated at a 95% probability level or not, but either way, it seems peculiar that it is referenced as a "general acceptable value" in reference to a *Repeatability* (*Limit*) *Requirement*. This seemingly implies that, simplistically, half the laboratories therefore failed to meet this requirement. No explicit reference to the data set is provided to explore this topic further.

Sub-Clause 9.3.3: Reproducibility

Reproducibility is defined in [68, Sec. 9.3] as: "related to the differences which are observed when the same oil sample is analysed by different laboratories (inter-laboratory reproducibility), or when it is analysed by the same laboratory over long periods of time (after several days, weeks or months) (intra-laboratory reproducibility)". *Inter-Laboratory Reproducibility* would potentially qualify for *Reproducibility* under the ISO 5725 series assuming the same method was applied [40, Sec. 3] and the laboratory was testing under *Repeatability* conditions [35, Sec. 6]. *Intra-Laboratory Reproducibility* is slightly different and is better described as a measure of *Intermediate Precision* and not *Reproducibility* [40, Sec. 0]. Part 3 of ISO 5725 is dedicated to *Intermediate Precision* [36]. For reference, this discrepancy cannot be explained by referring to the older version of the ISO series [94, Sec. 3].

It is stated in [68, Sec. 9.3] that the "inter-laboratory reproducibility has been evaluated by CIGRE as around ±20 % at medium concentration levels". Unlike the value of k given for *Repeatability*, this value is speculated to be an estimate of a *Limit* and without a *Coverage Factor*, (or $f = 1$). This is based on [66], where it seems the *Accuracies* were based on the same dataset. As was explained for Equation (18), the *Accuracy* uses either just the *Reproducibility* or also includes a method bias. This means that the *Reproducibility* should be equal to or less than the *Accuracy*—as was shown in Equation (19). Clearly, $0.20 > 0.15$ and so it is instead interpreted that the value given represents the *Uncertainty* between two samples, i.e. a *Limit*, $0.20 \approx \sqrt{2} \times 0.15$.

From ISO 5725's perspective on *Precision*, the statistical concept is as shown in Equation (37) [35, Sec. 5]:

$$y = m + B + e, \tag{37}$$

where, for a particular material tested, m is the general mean (expectation), B is the laboratory component of bias under *Repeatability* conditions, and e is the random error occurring in every measurement under *Repeatability* conditions. From these, Equations (38) and (39) give the estimated *Between-Laboratory* standard deviation, s_L, and estimated *Within-Laboratory* standard deviation, s_W, respectively. It is generally assumed in ISO 5725 that the estimated *Repeatability* standard deviation, s_r, is shared amongst laboratories following a *Method Procedure* such as [68]. Note that outliers are first removed which is a topic not explored in this thesis. Therefore, the mean value of

s_W is used, shown in Equation (40). The estimated *Reproducibility* standard deviation, s_R, is s_r and s_L combined, as shown in Equation (41).

$$s_L = \sqrt{var(B)}, \tag{38}$$

$$s_W = \sqrt{var(e)}, \tag{39}$$

$$s_r = \sqrt{\overline{var(e)}}, \tag{40}$$

$$s_R = \sqrt{{s_L}^2 + {s_r}^2}, \tag{41}$$

Simply put, this is the extent measurements may be expected to differ for a given time and for a given material across laboratories. There are nuanced relationships between s_r, s_L, and s_R, depending on the experimental design that is beyond the scope of the thesis and constitute the bulk of the ISO 5725 series. The values mentioned in [68, Sec. 9.3] likely refer to Equation (42) where the value for f is unknown.

$$0.2 = \sqrt{2} \times f \times s_R/m, \tag{42}$$

It is stated in [68, Sec. 9.3] that it is recommended for laboratories to check their own *Intra-Laboratory Reproducibility* but provides no indicative values as to what should be expected. It continues, stating that samples should be analysed "at regular intervals of time, for instance each week or each month over a period of several months" [68, Sec. 9.3]. As discussed, this is more a measure of *Intermediate Precision* [39, Sec. 6] which is explained further in [36]. This would correspond to the *Simplest Approach* described in [36, Sec. 6] for *Within-Laboratory Study*. In contrast to Equation (37), *Intermediate Precision* partitions B as shown in Equation (43) [36, Sec. 6]:

$$y = m + B_0 + B_{(\dots)} + B_{(n)} + e, \tag{43}$$

Within this model, B_0 is the "residual" component of the laboratory bias, and each other subscripted B are "effects corresponding to intermediate precision factors". Under *Repeatability*, these would all be an unknown constant forming B. Under the *Intermediate Precision*, only B_0 is considered a constant, and each of the partitioned components can then vary depending on what they represent. The estimated *Intermediate Precision* standard deviation, s_I, is as shown in Equation (44) [36, p. 27]:

$$s_I{}^2 = s_r{}^2 + \sum_{i=1}^{n} s_{(i)}{}^2, \tag{44}$$

where (i) corresponds to the *Intermediate Precision* factors, although note that B_0 is excluded. In this case, although time is the obvious factor being varied in the *Intra-*

Laboratory Reproducibility, it is intended to be *Confounded* with other factors not being controlled for, such as changes in *Calibration* or *Operator* [36, Sec. 6]. The relevance is that there should be an attempt made to ensure these factors **are** sufficiently varied during this period to obtain a representative estimate of *Intermediate Precision*. Many other experimental setups are discussed in [36] for further reference, and it advises at least 15 measurements where the factor, here: time, is varied between each measurement [36, Sec. 6]. For reference, [68, Sec. 9.3] states that samples should be analysed "at regular intervals of time, for instance each week or each month over a period of several months": it is not clear how measuring once a month for a period of several months would satisfy the ISO 5725 recommendations. Nevertheless, [36, Sec. 6] states that this *Simplest Approach* "can be useful for studying time-different intermediate precision by making successive measurements on the same sample on successive days, or for studying the effects of calibration between measurements". Much of ISO 5725 is on the minutiae of the experimental setup and the influences it will have on the correct formulation of equations for estimating parameters, therefore the above, especially regarding *Intermediate Precision*, should not be considered a comprehensive review.

Given that this thesis is assuming the CIGRE results discussed in [66, p. 22] are the source of the new guidance for *Inter-Laboratory Reproducibility* added in [91, Sec. 9.3] as "around ±20%", the other values in [66, p. 22] are also of interest. In particular, [66, p. 22] states that the average *Intra-Laboratory Reproducibility* of the CIGRE laboratories was estimated as ±10% at medium concentration levels. If no other information is available, and the default *Accuracy* of ±15% is being used, then this may serve as a default value for *Intra-Laboratory Reproducibility* for calculating *Uncertainty* in deltas or rates between samples—so long as they were analysed by the same laboratory. This is in line with the recommendations of [66, p. 23] with the caveat that if the measurements were also "made over the same day or a short period of time", then it suggests using *Repeatability*. Although, an important note of [66] is that it considers sampling as out of scope. Therefore, only the dates between the laboratory analyses would be relevant in its context. However, this was not done in Section 5.3 of this thesis. As the value for *Intra-Laboratory Reproducibility* was not included in [68, Sec. 9.3], it was assumed outside of the normative reference scope, and it was felt that insufficient information was provided explaining the origin of this estimated value to incorporate

it. For context, [66] cites [71] for the CIGRE data, which in turn cites [70] as a summary document, and then states the full version is available on e-CIGRE. Similarly, [70, p. 27] says the full version is available on e-CIGRE. However, this report could not be found for this thesis, even after making direct enquiries to e-CIGRE. When [1, Sec. 5] references this data, it cites [68] and [33]. As discussed, [68] does not cite the source of the data, and lastly, when [33, Sec. 10] references the ±10% for *Intra-Laboratory Reproducibility*, it cites back to [66].

Sub-Clause 9.3.4: Accuracy

Accuracy is defined in [68, Sec. 9.3] as "related to the differences that are observed between the values analysed by a laboratory and the true values…". It continues, stating the *Accuracies* should be determined using GIOS samples and gives "examples of accuracies that can be obtained using the overall experimental procedure … deduced from IEC and CIGRE inter-laboratory tests…" in its Table 6 [68, Sec. 9.3]. This Table is headed as "accuracy, in percentage of the nominal value" and captioned as "examples of accuracy of extraction methods". Thus, it is unclear how this is interpreted as a "required" *Accuracy*, nor how the commonly cited ±15% rising to ±30% at low concentration levels, get attributed to this source. The "source" is speculated to be from [66, p. 22] which tabulates average *Accuracies* for medium and low gas concentrations at ±15% and ±30% in its Table 2. The word "source" is in quotations because of the discussed ambiguity regarding the origin of these values.

Two points of corroboration are given here. First, when the 2003 publication [71] characterised the *Accuracies* in the older version [92, Sec. 9.3], it used the top row of values (±13% and ±35%) corresponding solely to the Toepler extraction procedure; indicating the change in characterisation occurred after that date. Second, the 2010 publication [33, Sec. 6] states: "the average accuracy of laboratories has been reported during IEC / CIGRE round robin tests to be around ±15% at routine gas levels", citing the 2005 [66]. As an aside, [66, p. 22] states to use the greater of either the *Relative Uncertainty* or a minimum *Absolute Uncertainty* equal to the LoD, *S*, included in [68, Sec. 9.2], when discusses the CIGRE average *Accuracy*. This is also perhaps why the phrasing was changed regarding *Repeatability* from Equation (26) to Equation (25) as mentioned earlier and corroborates the assumption that the value added for k was

intended as the value for b. The current phrasing of [68, Sec. 9.3] allows *Accuracy* to be smaller than *Repeatability* at very low concentration levels.

If evaluating via ISO 5725: [39, Sec. 4] states that if the laboratory component of bias, B, is unknown, Equation (12) is applicable to compare between measurements and a GIOS sample—assuming no *Uncertainty* associated with the latter. This means the *Critical Difference* would be as shown in Equation (45) [39, Sec. 4].

$$CD = \frac{1}{\sqrt{2}}\sqrt{\left(\sqrt{2} \times f \times \sigma_R\right)^2 - \left(\sqrt{2} \times f \times \sigma_r\right)^2 \left(\frac{n-1}{n}\right)}. \quad (45)$$

So, for either extreme of $n = 1$ or $n \ggg 1$ the *Critical Difference* would be as shown Equation (46) and Equation (47):

$$CD = f\sigma_R, \quad (46)$$

$$CD = f\sigma_L. \quad (47)$$

As with the rest of the values, the *Accuracy* values provided in Table 6 in [68, Sec. 9.3] have some associated ambiguity regarding their interpretation. Using similar logic as discussed, it is assumed that the equivalent of Equation (45), but for a single sample, is as shown in Equation (48):

$$A = f \times s_\delta / m, \quad (48)$$

where A represents the values given for *Accuracy* in Table 6 in [68, Sec. 9.3] and δ represents *Accuracy* in reference to $(\bar{y} - \mu_0)$. The *Coverage Factor,* f, is not known. Please note that the symbol A is used solely in this thesis, and not in ISO 5725 nor [68].

Example Impact of Ambiguity

It is difficult to source the original datasets to ascertain the intended interpretation of these values. This Sub-Section intends only to highlight the challenges with trying to deduce the source of these values and demonstrate significance of the issue.

One self-contained example is that the 1992 [92, p. 16] defines *Accuracy* as the "closeness of the true value to the mean of several measured values". Then, the 2005 version [91, p. 42] redefines *Accuracy* to "related to the differences which are observed between the values analysed by a laboratory and the true values…". The significance is that the definition in [96, p. 1], [92, p. 16] is now applicable to *Trueness* and **not** *Accuracy*. Yet, the specific values provided in the Tables for *Accuracy* did not change

once the definition was changed whereas it may be expected for a component of *Repeatability* to be added to the values.

In 1993, the unnumbered table in [91, p. 17] tabulated the *Accuracy* for 3 extraction procedures deduced from 19 laboratories. In 2005, this was updated to Table 6 in [91, Sec. 9.3], where it has results for 8 extraction procedures based on 44 laboratories. Annex C of [91], and now Annex C of [68], discusses most of the new extraction procedures, namely: mercury-free Toepler, mercury-free partial degassing, and the "shake test" methods. It states the example accuracies added in Table 6 in [68, Sec. 9.3] were based on an inter-laboratory test using 2 GIOS samples where 2 used mercury-free Toepler, 1 used mercury-free partial degassing, and 7 used the "shake test" method. In 2001, [70] discusses a CIGRE-led *Round-Robin Test* (RRT) that started in 1997 involving 25 laboratories. Adding 25 laboratories to the previous 19, gives the new 44 laboratories. Furthermore, the new extraction procedures that were the focus of [70], and the number of laboratories for each method, were the same as added to Table 6 in [91, Sec. 9.3]. Lastly, for every **new** extraction procedure, the values between Table 2 in [70, p. 198] and Table 6 in [91, Sec. 9.3] match—except for the "shake test" for reasons unknown. The existing extraction procedures remained unchanged however, despite the newly collected relevant data.

The 2003 publication, [71], states that it is "an intermediate, updated version of these documents", referring to [70] and the unlocatable full report. The number of laboratories and RRTs mentioned, and average *Accuracies* stated, remain unchanged between [71] and [70]. Therefore, this thesis assumes its description of the methodology is applicable. [71, pp. 8–9] states the following:
"Accuracies were calculated as the deviation, in absolute %, from the prepared values in the gas-in-oil standard, for each gas and each lab, using the results of RRTs 2A, 3b, 3A, 3B, and 4. The average accuracy for each lab was then calculated as the average of accuracies for each gas, excluding air. Finally, the average accuracy for each technique was calculated as the average of accuracies from each lab using this technique."

This description bears close resemblance to Equation (34). The significance is that it is using absolute differences, it is not combining in quadrature, and that it uses n rather than $n - 1$. The example in Annex E of [68], which is based on Appendix B of [33], can

be a useful point of reference. It first finds the bias of the laboratory by calculating the mean difference and then uses the mean absolute difference for estimating the dispersion about the mean. Thus, it can be inferred the *Accuracy* metric described in [71, pp. 8–9] is intended more as a metric of dispersion.

Some of the published data is utilised in this thesis to demonstrate the impact of selecting a given interpretation. Table AI and Table AII in [71, pp. 13–14] tabulate measured gas concentration levels for its **RRT 1** and **RRT 4**. **RRT 1** used samples taken from a decommissioned UK National Grid TX, meaning that the ground truth value was not known. It is stated in that [71, p. 8] the results were ultimately abandoned: "RRT 1 used oil samples taken from a transformer removed from service. The spread of results was such as to make it impossible to use the average of results as a reliable representation of the actual value". **RRT 4** instead used GIOS samples so that the expected reference values were known. Additionally, Table 12 from [33, Sec. 8] contains results of *Accuracies* from a presumed different RRT—called **RRT 409** henceforth. For simplicity, only the hydrocarbons are included in this thesis.

Several of these potential metrics are calculated and tabulated in Table 3-6. A metric called "spread" defined in Equation (49) is included, based on Table AI in [71, pp. 12–13] for **RRT 1**. The second metric is Equation (34) which is presumed similar to their *Accuracy* metric. The third metric represents a more typical estimate of a standard deviation, as per Equation (36). Lastly, Equation (10) and Equation (11) are included, representing an estimate of a uniform distribution and its estimated standard deviation. In the cases where an expected reference value was known, the metrics were calculated both relative to the arithmetic mean, and relative to said reference.

$$Spread = \left(\max_{i \in n}\{y_i\} - \min_{i \in n}\{y_i\}\right) \times \frac{100}{\bar{y}}, \tag{49}$$

Fig. 3-8 plots the data used in Table 3-6 for context, with the columns corresponding to **RRT 1**, **RRT 4**, **RRT 409**, respectively. The results show the measured gas values relative to the mean of their respective population. The points' colours and shapes correspond to a given gas. Where an expected reference value was present, it is plotted as a coloured vertical line. The horizontal lines separate the different extraction procedures used by the laboratories and are labelled accordingly. The data points for each gas are then used to estimate an empirical distribution, implemented using the

default parameters of the default "density" function in R along the top. Preliminary inspection would indicate a bias in the method(s) towards underestimating gas values given that the averages for all gases were less than the expected reference values. Furthermore, there is a greater relative variability in **RRT 1** and **RTT 4** [71], than in **RRT 490** [33]. This would prompt further scrutiny under the ISO 5725 procedure [35, Sec. 8], [89, Sec. 7].

Table 3-6: Accuracy of Laboratories using Gas Extraction Methods

Method / Data	%, Relative to Mean				%, Relative to Reference Value			
Data: RRT 1[a]	CH_4	C_2H_6	C_2H_4	C_2H_2	CH_4	C_2H_6	C_2H_4	C_2H_2
Equation (49)	77	75	76	83	–	–	–	–
Equation (34)	17	18	16	22	–	–	–	–
Equation (36)	21	23	21	27	–	–	–	–
Equation (10)	56	54	54	50	–	–	–	–
Equation (11)	33	31	31	29	–	–	–	–
Data: RRT 4[a]	CH_4	C_2H_6	C_2H_4	C_2H_2	CH_4	C_2H_6	C_2H_4	C_2H_2
Equation (49)	85	129	87	71	75	123	86	65
Equation (34)	11	14	12	17	14	16	13	15
Equation (36)	15	24	18	20	18	23	18	20
Equation (10)	47	103	64	43	54	94	62	48
Equation (11)	27	59	37	25	31	54	36	28
Data: RRT 409[b]	CH_4	C_2H_6	C_2H_4	C_2H_2	CH_4	C_2H_6	C_2H_4	C_2H_2
Equation (49)	55	61	85	55	52	56	82	50
Equation (34)	9	11	13	10	10	13	13	11
Equation (36)	14	15	19	14	14	17	19	16
Equation (10)	38	36	62	35	42	43	57	42
Equation (11)	22	21	36	20	24	25	33	24

Source Data: a): Table AI and Table AII in [71, pp. 12–13], b): Table 12 from [33, Sec. 8]

It is therefore not clear where in the ISO 5725 series the intended use of this metric is explained, or what these values represent. For example, [39, Sec. 4] considers the case of comparing measurements from multiple laboratories against a *Reference Material*, and it uses the mean difference and not the mean absolute difference. Arguably, [68] should highlight when values or definitions deviate from ISO 5725 to avoid ambiguity. It should be noted that even if the intended interpretation of the given metrics were known, there is still a significant *Uncertainty* regarding the estimate of the parameters given how few samples were typically used. Annex 1 of [35] discusses the *Uncertainty* in the estimates of these parameters relative to the number of required laboratories and tests. As per [66, p. 23]:

"Because of the economic and practical realities of laboratory DGA, the usual practice ... is to base their measurement accuracy estimates on the average error of only one or two measurements of gas-in-oil standards. ... it must be recognized that when these accuracy

figures are used for basic statistical inference, the statistical significance level is unknown. This caveat applies to the CIGRE and IEC results quoted as well as to individual lab accuracy estimates."

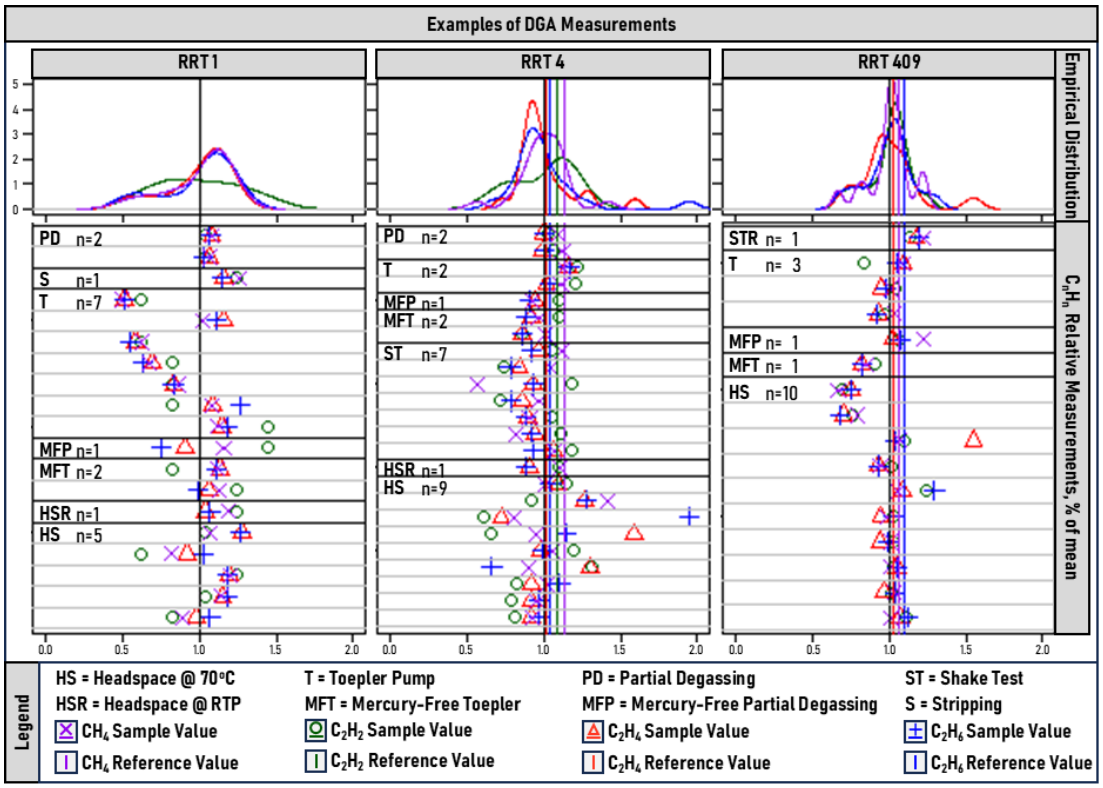


Fig. 3-8: Accuracy of Laboratories using Gas Extraction Methods

3.3. Conclusion

This Chapter contributes towards addressing *Research Theme 1A* by providing a detailed review and conceptual comparison of the four chosen TX DGA methodologies: IEEE C57.104-2019 [1], IEC 60599:2022 [2], NEI [1, Annex F], [3], [4], and LSA [5]. These will be used in Chapter 4 to develop automated implementations to provide practical comparisons. This Chapter argues that the IEEE's *DGA Status* levels of 1–3 do not align with the IEC's *Typical*, *Alert*, and *Alarm* scale. **Ł1** aligns best with *Typical*, **Ł2–3** lie closer to *Alert*, with the IEEE's *Extreme DGA Results* aligning best with *Alarm*. It also highlights that the IEEE methodology is conceptually focussed on solely providing a *Screening* output rather being indicative of *Fault Severity*. This narrower scope is aligned with literature in CIGRE on appropriate *Transformer Assessment Index* (TAI) design discussed in Chapter 2.

This Chapter also contributes towards addressing *Research Theme 2* by providing a detailed Standards-based literature on the topic of *Uncertainty* as applicable to IEEE C57.104-2019. The Chapter concludes that the inclusion of *Uncertainty* is a nuanced and complex topic, with no singular pre-determined approach viable for all scenarios. The Chapter emphasises how *Measurement Uncertainty* estimates may overlook other critical aspects of *Uncertainty*. Furthermore, the challenges with interpreting some of the relevant metrics provided for *Measurement Uncertainty* for DGA was discussed in detail, with particular focus on IEC 60567:2011 [68].

Although this thesis chose a *Gaussian* ($\mathcal{N}$) distribution of ±15–30% at a 95% probability level for *Accuracy* as the only metric, some alternatives were presented. As the general statistical models used for *Measurement Uncertainty* allow for partitioning of components, they can similarly allow new components to be included. However, it was highlighted that the practical application of such a methodology would face challenges due to the specific derivation of the output metrics in IEEE C57.104-2019. Attempting to assert both *Accuracy* and *Reproducibility* measures simultaneously will lead to instances of incoherent assumptions that cannot be met.

4.Comparative Analysis of Methodologies

Chapter Purpose

This Chapter presents analyses and findings relevant to the practical deployment of the reviewed DGA methodologies: IEEE C57.104-2019 [1], IEC 60599:2022 [2], NEI as outlined in [1, Annex F], and the LSA methodology [5]. It can be difficult to predict how methodologies behave comparatively prior to investing time to implement them. Therefore, it is informative to a would-be user of a given methodology to have access to real TX DGA case studies to explore potential outputs and comparative behaviours. Furthermore, the findings in this Chapter can expediate the deployment process by pre-emptively highlighting potential barriers to a practical implementation. Lastly, as *Uncertainty* is a broad and complex topic, there are potentially many relevant factors to consider. The findings in this Chapter can help justify the inclusion and/or exclusion of certain considerations to define a more manageable scope.

Chapter 3 provided conceptual comparisons of the reviewed methodologies whereas Chapter 4 provides practical comparisons using simple models in the preliminary analysis, and real TX DGA data in the case study analysis. Therefore, *Research Theme 1A* is concluded here as the implications of the changes made to the new methodology outlined in IEEE C57.104-2019 have been addressed. Lastly, the findings from this Chapter contribute to *Research Theme 1B* by motivating and justifying the improvements proposed for IEEE C57.104-2019 in Section 5.1.

Chapter Structure

Section 4.1 conducts basic conceptual experiments to quantify the potential impacts of various factors using real TX DGA data and limits from IEEE C57.104-2019. A major difference between the methodologies outlined in the IEEE and the IEC is the former's use of a linear regression model to estimate gassing rates. It claims it substantially reduces *Uncertainty* but adds various stipulations regarding the number of samples to be used. Some aspects, such as why [1] might consider 6 samples an upper bound for its linear regression model is explored here. The intent is gaining an understanding of the potential impacts of said decisions, rather than a critique of any decisions made.

Section 4.2 provides a background how each reviewed methodology was interpreted for automated implementation for this thesis. These implementations represent the

'base' form attempting to adhere to the original guidance as close as possible. The relative behaviours of these implementations are investigated using some historic TX data as case studies. The focus is on identifying practical barriers to deployment, including those introduced by the changes made to the IEEE C57.104 methodology. These results are discussed and motivate proposed improvements introduced in Chapter 5 for IEEE C57.104-2019 as part of *Research Theme 1B*.

4.1. Preliminary Analysis on Measurement Uncertainty

4.1.1. Motivation

This Section explores factors that influence *Uncertainty* and their quantification. This is distinct from the previously discussed topic of interpreting the *IEC Specification*. The focus is on exploring the choice of metrics made by IEEE C57.104-2019, which are asserted in this thesis to be motivated by mitigating *Uncertainty* issues intrinsically.

The change in gas level metric used in 𝕋3 is "dominated principally by the DGA result fluctuations caused by the analysis process itself" [1, Sec. 6.1], and is intended to distinguish normal variations from other causes. One potential perspective to consider is through the lens of *Reproducibility*. If two samples are outside of the *Reproducibility Limit*, then they can be considered for further attention. Considering that the primary consequence of failing 𝕋3 is to obtain a *Confirmation Sample*, it echoes the ISO 5725 series' sentiment on *Conformity*. It is stated in [39, Sec. 5] that "if there is any suspicion that the test result may not be correct, a second test result should be obtained", and it then suggests two approaches for utilising the new results depending on their difficulty to obtain. Generally, if the new results are within the *Reproducibility Limit*, they are combined, otherwise the more recent results are kept. In [1], the difference is that either only the most recent result is kept, or both results are kept, without combining.

The metric of change in gas levels as normalised over a year using multiple samples, as used in 𝕋4, attempts to mitigate *Measurement Uncertainty*. As per [1, Sec. 6.1], "as the number of points increase, fluctuations caused by the laboratory DGA analysis process cancel each other (average out)". Furthermore, as was discussed, and will be further demonstrated, the metric of average gassing rate is less susceptible to inhomogeneity than a pairwise comparison. Lastly, obtaining a *Confirmation Sample* once something

unusual is detected can help mitigate the non-repeatable components of all three sources. These measures combined aim to reduce unnecessary flagging from *Screening*.

In contrast, IEC 60599:2022 does not appear to consider reducing *Uncertainty* specifically in relation to its chosen metrics. The two tangential topics were the use of changes in levels to calculate ratios where appropriate, and a warning in [2. App. A.2.5] to not attempt to use gassing rates to project future gas concentration levels to then compare against tables such as Table A.2 as factors such gas losses would be neglected.

4.1.2. Scope

This Section provides a preliminary analysis of some common assumptions rather than definitive best practice recommendations. Various impactful topics are being neglected here. As was mentioned, only the *Uncertainty* related to the measurement and the metric is considered: considerations of the limit and *Diagnosis* are considered out of scope. Since the thesis's focus is on *Screening* as opposed to *Diagnosis*, the focus will be on absolute gas levels and changes in gas levels as opposed to the ratios of gas levels.

Gases are dissolved in a shared medium and the units of parts per million (ppm) imply an interaction between the gases. This can be reframed by asking to what extent does a high presence of one dissolved gas in the oil influence the measurements of other gases. An increase in one gas does not simply displace or dilute the concentration of another gas. The interactions depend on factors such as oil type, liquid saturation, temperature, pressure, gas solubility, and potential interactions between gases. In addition, there cannot be less than zero of a given gas. Assuming a fixed distribution shape may result in confidence intervals crossing the zero threshold. It is unclear the extent this is relevant in many applications. This topic is in [32, Sec. 5], where it lists potential transformations for further reference. Generally, for univariate cases, having a *Relative Uncertainties* $\leq$ 100%, or an *Absolute Uncertainty* $\leq$ LoD, has negligible impact. For multivariate cases that have estimated correlations, negative values may more frequently be naïvely expected. These topics are not considered further here.

Correlations between variables is a significant complication in calculating the probability of *Non-Conformity*. For example, if loading affects gases within a TX, then presumably all gases would be affected to some extent. Calculating the estimated likelihoods for each gas separately, assuming independence prior to combining, can

give misleading outputs. It has historically been difficult to quantify these relationships, especially when taking say one sample a year. Perhaps now with the increasing prevalence of high sampling OLDGA, these relationships may become better quantified, with guidelines established. Nevertheless, the remainder of this Section assumes gases are independent, although this is briefly revisited in Section 5.2.

Measured Values versus True Values

The following topic is only relevant where *Relative Uncertainty* is assumed, such as with the *IEC Specification*. As discussed, a symmetric distribution such as a $\mathcal{N}$ distribution is often assumed. If it is applied to a measured value, it can imply a bias due to the asymmetry caused when considering different measured magnitudes. This can be intuited thusly: if the true value, μ, lay at the lower end of the measured estimate, $\hat{y}_1$, the error, δ_1, would be greater than if true value lay at the upper end of a different estimate, $\hat{y}_2$, as shown in Equation (50). This would imply a tendency for the true value to be higher than the estimate. Arguably, the distribution should be applied to the true value, which of course, is unknown. In practice, if multiple samples are considered, ISO 5725 uses the mean of the samples, m.

$$(\delta_1 > \delta_2)|(\mu = \hat{y}_1 \times 0.85, \mu = \hat{y}_2 \times 1.15). \quad (50)$$

The plot in the left of Fig. 4-1 has along the abscissa the 'true' gas level. Its *Measurement Uncertainty* is contoured along the ordinate based on an assumed $\mathcal{N}$ distribution of ±15% assumed at a 95% probability level. It shows as the 'true' gas level rises, so does the spread of the potentially measured gas level. For the sample value of 50 ppm, this is projected to the side plot, showing the expected normal distribution. Projected upwards to the top plot is the probability distribution of the true gas value given the measured gas value of 50 ppm. The highlighted region in red is the bias introduced via the shift in the expected value. This is further demonstrated in the bottom right plot of Fig. 4-1 which explores the generated distributions given different *Relative Accuracies*, ranging from ±10% to ±30% at a 95% probability level. The results similarly show a clear skew and shift in the expected value.

The results are based on a *Monte Carlo Method* (MCM) approach where a total of 20^7 samples were drawn at set intervals across a range, where each interval represented a true value from which a normal distribution was calculated. Then, from each interval's randomly generated values, the count of the desired measured sample was retained.

These counts, which summed to approximately 10^6 samples, were used to create a probability density. There is feasibly an analytical solution to describing the distribution not pursued given how in the top right plot of Fig. 4-1, the distributions appear geometrically similar for 25 ppm in black and 50 ppm in red for ±30%.

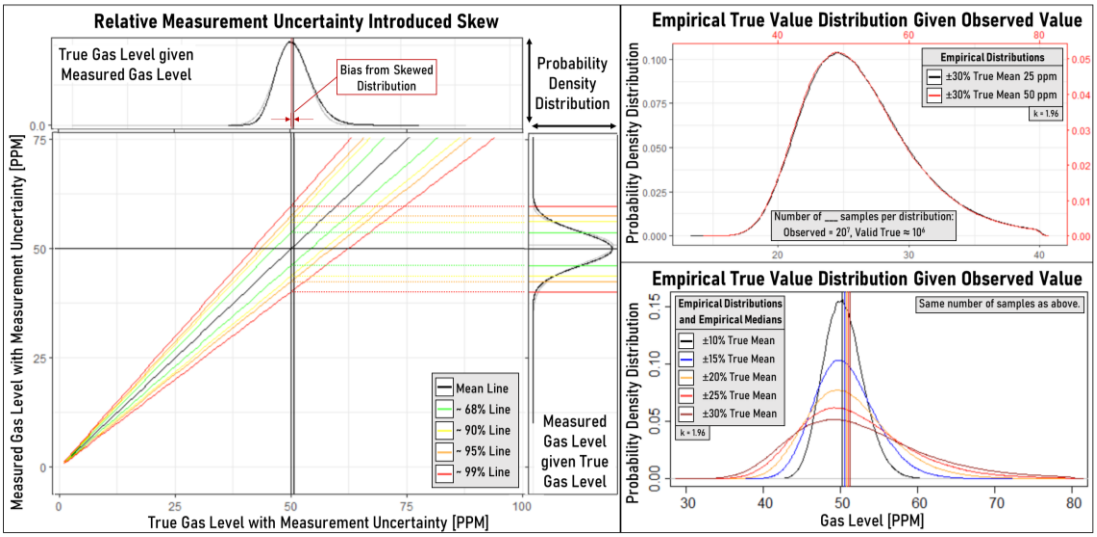


Fig. 4-1: Comparing Uncertainty in Measured Value and in True Value

For estimating gassing rates via a linear regression model, Fig. 4-2 compares the impact of not accounting for the bias using example TX DGA data, which will be further explored and explained later. The gases were subsampled down from 480 samples to 48 samples, taken at equidistant points, to reduce the computational burden. Black is used to denote the scenario where the predicted 'true' values are used which are estimated using the same MC-like approach as described for Fig. 4-1. The standard deviations of the estimated parameters are based upon on the empirical distribution obtained. Red is used to denote the scenario where the observed values are used, and the standard deviations of the estimated parameters are based upon the using a *Fixed-Effect* (FE) model where the ±15% *Relative Uncertainty* is input explicitly. The FE model will be explained in more detail later in this section. The results indicate that although the intercept point can noticeably change, the impact on the slope coefficient is largely insignificant as the biases mostly cancel one another out. These are generally assumed an artefact of the simplified model used to represent the *Uncertainty* rather than a true phenomenon. For example, the biases shown in Fig. 4-1 are not expected to be related. Therefore, this topic is not explored further.

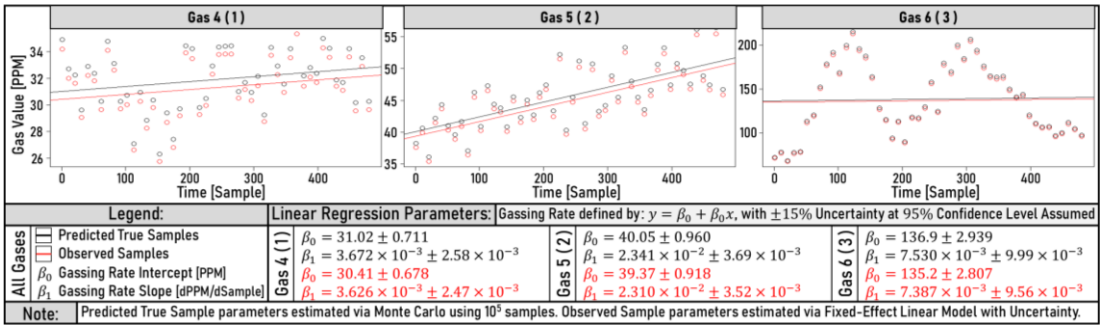


Fig. 4-2: Comparing Gassing Rate between Observed Samples and Predicted True Samples

4.1.3. Absolute Gas Levels

The simplest metric is the absolute gas concentration level. The focus here will be on *Relative Uncertainty* assuming a *Gaussian* ($\mathcal{N}$) distribution at a 95% probability level. The aspects explore are the magnitude of the *Relative Uncertainty*, the inclusion of LoD-based interpolation of said magnitude, and the inclusion of a minimum *Absolute Uncertainty* as suggested by [66]. Given that in the context of absolute gas levels, it is the upper limit that is most relevant for *Condition Assessment*, the limits within [1, Sec. 6] will be used as points of reference when considering the impact.

Fixed Relative Uncertainty

Relative Uncertainty results in greater *Absolute Uncertainty* nearer and above the limits than at lower values. The plots in Fig. 4-3(a) explore the impact of the two extremes in the *IEC Specifications*: ±15% and ±30%. This *Uncertainty* is applied to C_2H_4 centred on the limit of Table 1 of [1, Sec. 6] for a sealed TX of unknown age; 50 ppm. The ordinate shows the likelihood of a sample being mislabelled: i.e., a false positive or negative. Values below the limit are shown in purple, and those above in orange. The latter are reflected backwards on the x-axis to better compare the differences. The grey line represents applying the *Uncertainty* to the limit as opposed to the sample value. As this is constant, it represents a natural point of comparison. The results show that higher gas values are more likely to be mislabelled, which may be problematic as this represents a failing value being erroneously classed as passing.

Scaling Relative Uncertainty

The plots in Fig. 4-3(b) explore the impact of *Limit of Detection* (LoD), (S). The *IEC specification* is typically cited as having ±15% at medium concentration levels and ±30% at low concentration levels. Its Table 6 uses for low concentration levels, values

between 1–10 ppm and 30–100 ppm for hydrocarbons and carbon oxides, respectively [68, Sec. 9.3]. A threshold of 10 ppm is given in [68, Sec. 9.3] to differentiate *Repeatability* between low and medium levels. [2, Sec. 6.2] states that the *Uncertainty* is typically ±15% at $10 \times S$, and that it increases rapidly to typically ±30% at $5 \times S$. Although the method to interpolate values within this range is not specified. Therefore, Equation (56) is assumed as the relationship used in Fig. 4-3(b), where U is the *Relative Uncertainty* and y is the gas level of the sample:

$$U = \begin{cases} 0.30 & G \leq 5 \times S, \\ 1.50 \times \dfrac{S}{y} & 5 \times S < y < 10 \times S, \\ 0.15 & G \geq 10 \times S, \end{cases} \tag{51}$$

where [68, Sec. 9.2] gives an LoD of 1 ppm for hydrocarbons for example. This gives a slight non-linear relationship to accommodate the "increases rapidly" description. As discussed before, [35, Sec. 8] also outlines other candidate functional relationships.

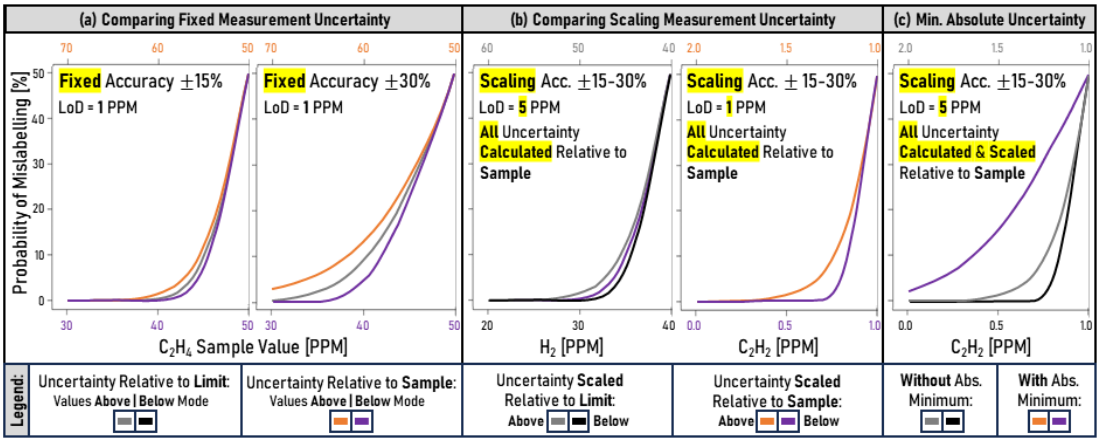


Fig. 4-3: Comparing Measurement Uncertainty Approximations

To gauge the relevance of *Relative Uncertainty*, the LoDs for different gases are compared in reference to their limits, as it is argued quantities of a gas much lower than its limits is of less concern. In these examples, LoD values are according to Table 5 of [68, Sec. 9], and limits according to Tables 1 and 2 of [21, Sec. 6]. Two gases have LoDs near enough to their respective limits where this topic may be relevant. Hydrocarbons in general have an LoD of 1 ppm, but only C_2H_2 has a low enough limit, which ranges between 1–7 ppm depending in the specific details. The second gas is H_2, which has an LoD of 5 ppm and a limit ranging between 40–200 ppm. As the limits for H_2 and C_2H_2 are at $8 \times S$ and $1 \times S$, respectively, both would clearly have a greater degree of

Uncertainty if using an increased relative *Uncertainty* at low gas values—with C_2H_2 being impacted more.

To analyse the impact of interpolating across different magnitude *Relative Uncertainties*, the black and grey lines in Fig. 4-3(b) calculate the **scaling** of the magnitude of the *Relative Uncertainty* relative to the limit whereas the purple and orange lines calculate it relative to the sample value. To emphasise, in both cases the *Uncertainty* is calculated relative to the sample value. C_2H_2, shown on the rightmost plot of Fig. 4-3(b), exhibits no difference in this gas range via either approach as both will be capped at the maximum ±30% given the low values. However, this would not be the case for other candidate limits, as for example, a free-breathing TX of unknown age in Table 2 of [21, Sec. 6] has a limit of 7 ppm: above $5 \times S$. In contrast, H_2 shows an interesting result where the baseline of applying the scaling of the magnitude of the *Uncertainty* relative to the limit saw a greater asymmetry than applying it to the sample. However, this is driven by Equation (56); a different interpolation relationship would perhaps not have the effects of reducing absolute gas levels whilst increasing *Relative Uncertainty* cancel out so evenly. Although it is not clear which affect would be stronger given a different interpolation relationship, they can be expected to reside within the envelope created by scaling relative to the limit.

Minimum Absolute Uncertainty

There seems no basis to have the *Relative Uncertainty* increase as values approach $5 \times S$, but then remain constant below that. For example, [68, Sec. 9.3] uses S as the lower bound for *Repeatability*. It is instead suggested to pick the larger of a given *Relative Uncertainty* and an *Absolute Uncertainty* in [66]. An alternative is to have a constant to add to the scaling component. This was discussed for Equation (28) for *Repeatability* from [35, Sec. 8], where it was explained the older version of the *IEC Specification* [92, Sec. 9] used a similar approach, shown in Equation (24). The values suggested by [66] for the *Absolute Uncertainty* matched the LoD values in Table 5 of [68, Sec. 9] which would be equivalent to ±20% at $5 \times S$. Fig. 4-3(c) highlights the potentially significant difference in using such a minimum *Absolute Uncertainty* for C_2H_2. However, in contrast, no other gas saw a difference near their limits.

4.1.4. Relative Gas Levels (Deltas)

Tracking the gassing rate is often considered more important than referring solely to the absolute gas concentration values as it is thought to indicate current events rather than an accumulation of all historic events. However, the implications of *Uncertainty* differ. The focus here will be on the differences between the change in levels or delta, normalised delta, and use of linear regression for estimating the gassing rates are considered.

Choice of Uncertainty Metric

One of the issues with relying solely on metrics such as *Reproducibility* is it assumes consistent performance whereas factors such as sensor drift are known to occur over time. Relying solely on a static *Accuracy* evaluation, as is done in this thesis, will instead tend to overstate the *Uncertainty* unless correlations are factored in. Regardless of the distribution used, the *Relative Uncertainty* of the measured delta can be very high as it is dependent on the two absolute values and not on the estimated delta. If assuming independent $\mathcal{N}$ distributions, the delta would have a mean value as expected with a standard deviation that is equal to the square root of the summed variances. A summary of this kind of arithmetic for independent $\mathcal{N}$ distributions is outlined further in Table 2 of [97, p. 58]. As the variance is linked to the absolute values due to the *Relative Uncertainty*, larger absolute values result in a larger *Uncertainty* in the calculated delta. This means that unlike for *Absolute Uncertainty*, no approximation based on the limit would be valid for *Relative Uncertainty*.

Estimating Measurement Uncertainty

For *On-line DGA* (OLDGA), sufficient data typically exists to estimate short-term *Precision* but not *Accuracy*. There are complicating factors such as potential *Calibration* drift that prevent this that are considered out of scope. For further reference, [32, Sec. 10] discusses the topic of drift. To estimate the *Precision*, it is easier in cases where the gas can be considered 'constant' such that the fluctuations can be attributed to either variability or 'inhomogeneity' and/or *Measurement Precision*. If the gassing rate per sample is compared to the gassing rate of the entire duration, the differences can be considered as inhomogeneity. In this context, if the sampling interval is sufficiently frequent and the variability appears random with no autocorrelation, then the

distinction between variability and *Measurement Precision* is arguably irrelevant. Although, care must be taken in case its behaviour changes under new circumstances.

Consider Fig. 4-4, where three examples are of absolute gas values are plotted over time. **Gas 1** is assumed constant, and a histogram of its values is plotted beneath with the highlighted sample being considered an *Outlier* and thus ignored. Based on this, a $\mathcal{N}$ distribution can be fitted with an estimated 95% probability level of ±1.6 ppm which would be approximately ±5% *Relative Uncertainty*. This is done by first estimating the fitted $\mathcal{N}$ distribution via Equations (8) and (9), and then multiplying the standard deviation shown in Fig. 4-4 by a *Coverage Factor* of $k = 1.96$. In contrast, **Gas 2** cannot be assumed constant and must first be detrended. In this example, the initial 48 samples are ignored, and a linear trend is calculated for the remaining samples. The histogram shows the detrended values with a superimposed fitted $\mathcal{N}$ distribution that has an estimated interval of ±1.8 ppm. Care should be taken if attempting to approximate it to a *Relative Precision*, naïvely estimated at ±4.5%. If there were heteroskedasticity due to the positive trend, it would be lost in the transformation to a $\mathcal{N}$ distribution.

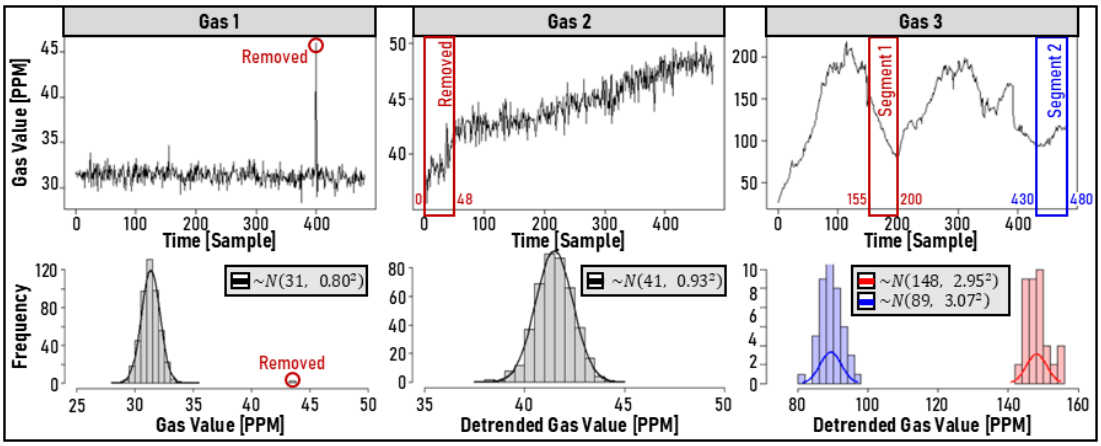


Fig. 4-4: Estimating Measurement Uncertainty from Samples

Lastly, **Gas 3** has clear variability that is distinct in behaviour from typical *Measurement Precision*. In such cases, it can be difficult to distinguish the trend from the noise. One crude method is to isolate segments that seem reasonable approximations of linearity. In this case, two segments are highlighted in red and blue, respectively. If these were detrended and fitted to a $\mathcal{N}$ distribution, their estimated intervals would be ±5.8 ppm and ±6.0 ppm, respectively. If they were similarly estimated as *Relative Uncertainties*, they would be naïvely estimated as ±3.9% and ±6.7%, respectively. In this case, it would

seem more likely that the *Absolute Uncertainties* are the more reasonable assumption rather than the *Relative Uncertainties*. Another point is that the residuals are not normally distributed even after detrending, especially for the first segment in red, thus invalidating the results to some extent regardless. Therefore, care should be taken when estimating parameters.

In Sub-Section 5.2.5, instead of a linear detrend, a moving median filter is used to filter out some variability. However, the efficacy depends on the parameters selected such as the window size and it is generally only effective for simpler cases or as an estimate. As explained in Appendix B of [89], the *ISO 5725* series is generally based on ANOVA, and this could potentially be used to attempt to partition different factors more effectively. ARIMA models is suggested in [32, Sec. 11] for time-series data and would likely be the most robust approach. Although, it is not clear how applicable they are in the context of as few as 4–6 samples. Unfortunately, neither are explored further in this thesis.

Uncertainty in Gassing Rates

One issue with using the metric of changes in gas levels directly is that they are a function of time and so potentially closely linked to the sampling interval. Therefore, when trying to evaluate an acceptable delta, some may choose to normalise to a set interval to obtain a gassing rate that is ostensibly more comparable. However, as the considered interval is shortened, the calculated normalised gassing delta tends to increase in variance [1, Annex B.1]. For intuition, if it is assumed two components to the data: the 'true' gas levels and resulting gassing rate, and an assumed *Uncertainty*. Any projections extrapolating will amplify both the *Uncertainty* component and the gas delta similarly. This will lead to over-estimations in the magnitude of the deltas and thus gassing rates. Over a given interval, true gassing rates will cause an accumulation of gas levels whereas the *Uncertainty* will not similarly accumulate.

This can be shown simply as in Fig. 4-5. Using real gas values from the same three examples in Fig. 4-4, the delta between the given sample and the initial sample, scaled to the overall duration are plotted in red. The region shaded indicates the 95% probability level assuming *Relative Uncertainty* of ±15%. In all three cases, the calculated gassing rates tend to reduce in volatility as the interval is increased. This highlights the importance of the sampling interval and how it is not as simple as

assuming a shorter interval leads to more accurate results. Furthermore, when comparing gassing rates, one should also be wary of assuming normalising the rate to a given fixed interval can provide a comparable metric. Rescaling the gassing rate as calculated during a one-hour interval to an annual rate will almost certainly differ drastically to a gassing rate calculated directly from a one-year interval. As per [1, Annex B.1], "normalization to a common time interval (year) simply does not work".

Given that multiplying a $\mathcal{N}$ distribution by a scalar value will result in another $\mathcal{N}$ distribution, one could simply rescale the calculated *Uncertainty* for the delta. This can also be used to gauge whether the calculated gassing rate is significant relative to the *Uncertainty*. Assuming there is gassing, at either a great enough interval or gassing rate, the increase should be significantly discernible from the *Uncertainty*. For example, it can be estimated with an approximate 95% confidence that there is a positive trend after a 300-sample interval for **Gas 2** in Fig. 4-5.

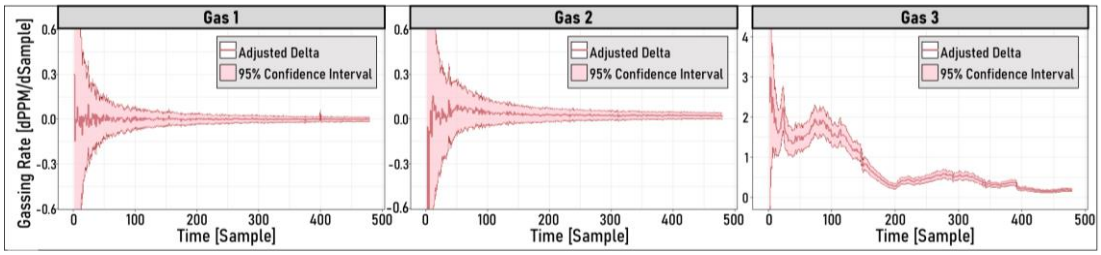


Fig. 4-5: Impact of Sample Interval on Gassing Rate Uncertainty

Use of Linear Regression

Using multiple samples could improve the result's reliability. In [66, p. 23], linear regressions are suggested: "even with a series of only three or four measurements, the accuracy of the estimated rate of increase will generally be much better than the accuracy of a rate calculated from a difference". The most recent version of [1] also employs this approach to "reduce the impact of these intrinsic DGA process variations on the rate determination" on the assumption that "the variations caused by the DGA process are random" and so "they will tend to cancel each other" [1, Annex B.1].

Although the premise is intuitive, it is challenging to qualify. The key issue is regarding intent: if the metric is intended to encapsulate the accumulation of gases over the given timeframe, then the start and end points are of primary relevance. A linear regression instead leverages interim points to create a more robust estimate of the average gassing rate. There is a topic regarding *Measurement Uncertainty* of the metric, and another

more fundamental topic of the relevance of said metric to its intended purpose, e.g. CMA. For example, [33, Sec. 10] states that "caution should be used when looking at gassing rates calculated over short periods of time, as these may be due to normal variations of gassing with operating conditions (e.g. load, ambient temperature) rather than actual fault". In that case, even if there is little *Uncertainty* regarding the correct estimate of the regression line, it may no longer represent the metric of interest.

Nevertheless, the concept of using an average to reduce *Uncertainty* is sound. For example, [90, Sec. 5] states that if systematic inhomogeneities are present, an average of measurements made over the parameter of interest may be used. In the context of gassing rates, the average rate or slope over the duration would be used. However, a linear regression is of course assuming a linear trend which may not be representative of the data. Aspects such heteroscedasticity and serial correlation all undermine the validity of an estimated linear trend. Therefore, care should be taken to not presume the metric faithfully represents the gassing rate specifically, but rather a simple method to produce a readily comparable metric related to the average gassing rate.

The fundamental difference between a delta-based approach and a linear regression is highlighted in Fig. 4-6. Using a Heaviside step function, $H(x)$, that triggers at point x', it plots the predicted gassing rate bounded by 95% confidence intervals using ±15% *Uncertainty*. The delta-based approach is in red whereas the linear regression is in green. It highlights that if the intention is to encapsulate the gassing over the interim period, the linear regression is sensitive to **when** the gassing occurred in a manner that a delta-based approach is not. Arguably, human analysis would also be sensitive to timings. If only a single sample at the either extremity shows a change from the interim period, perhaps one would assume noise until confirmed by a subsequent sample(s). Additionally, a step change nearer the end of the interim implies greater *Uncertainty* regarding its future performance than one occurring earlier as the subsequent stability has been maintained and demonstrated for longer. However, neither of these aspects are directly captured via the approaches. Gassing rate is arguably primarily used as a proxy for *Fault Severity* in retrospective analysis rather than a forecasting tool. It is a measure of gassing activity over the interval. As per [1, Sec. 1]:

"It should also be noted that DGA is a detection and diagnostic tool, not a predictive technique. It can only detect an existing or past condition and has no capability to

"predict" any future condition. However, when a condition exists, DGA can be used to track and evaluate its evolution over time".

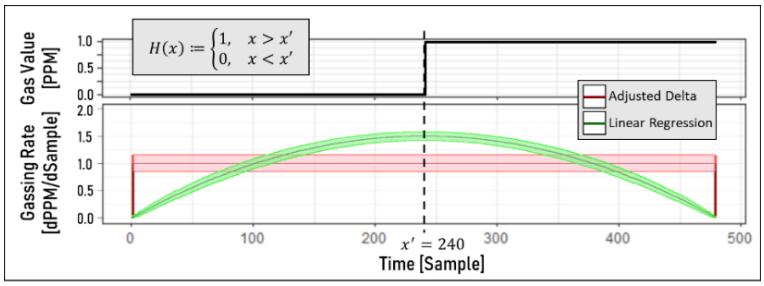


Fig. 4-6: Comparing Delta-Based Gassing Rate to Linear Regression using step-function

There is of course nuance to this. It is commonly accepted that in *Condition Assessment* via DGA, ceteris paribus, priority should be given to accelerating rates of changes, over rates of changes, over absolute values. There is therefore an element of implicit forecasting within the *Condition Assessment* process to 'evaluate its evolution over time'. Arguably, this aspect should be tackled explicitly as it is otherwise not clear to what extent one aspect should be prioritised over another. For example, how rapidly should a low gassing rate be accelerating to be considered a higher priority than a stable but high gassing rate. This remains an unresolved matter.

Linear Regression Uncertainty

Returning to the original premise that linear regression provides more accurate and reliable gassing rates; this can be interpreted in two ways. The first is in reference to the *Uncertainty* of the estimated gassing rate, and the second is the inherent volatility of the estimated gassing rate accounting for factors such as variability. Fig. 4-7, using the same gas examples, compares gassing rates estimated via normalised deltas, labelled "Adjusted Delta", to those from linear regression. The bottom row shows the estimated gassing rate based on all samples between the given sample and the initial sample via linear regression in green, and the delta-based approach in red. The top row has gas values in black with the trend line estimated via linear regression of all samples in blue.

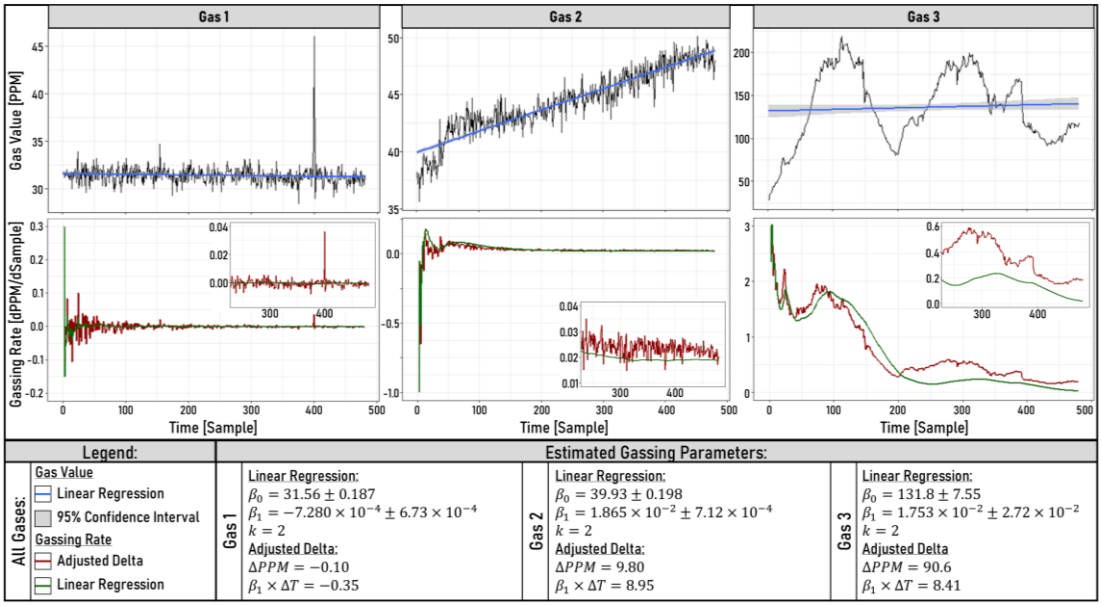


Fig. 4-7: Comparing Delta-Based Gassing Rate to Linear Regression

The results indicate that a decreased variance in the estimated gassing rate and a faster convergence to the apparent gassing rate, supporting the use of linear regression. Furthermore, when looking at **Gas 1**, it demonstrates a resilience to noise, where the outlier value at approximately sample 400 caused a noticeable change in the gassing rate via the delta-based approach but not via the linear regression-based approach. However, **Gas 2** arguably highlights how this very behaviour may be problematic where its apparent step-change due to gassing at approximately sample 50 is, roughly speaking, 'ignored' by the linear regression in its inherent assumption of a single linear trend. This is shown by the blue line in Fig. 4-7. This is undesirable if the intent is to characterise the change in gas levels in the duration, normalised to a given interval. Similarly, it would be very difficult to reconcile the gassing rate obtained via the linear regression for **Gas 3** as for the delta-based approach.

To better understand this behaviour, it may be useful to differentiate the delta-based approach, with its single source of *Uncertainty* via the *Measurement Uncertainty*, from the linear regression that could be considered to have two forms of *Uncertainty*. One being related to the *Uncertainty* represented by the confidence interval of the coefficient parameters, and the other being the overall *Uncertainty* including the variability of the residuals about the trend line, represented by the prediction interval. As an aside, where the thesis previously used a caret to represent an estimate, a tilde will be used henceforth: a caret will represent a maximum value, and it inverted will

represent a minimum value. Assuming linear regression via *Ordinary Least Squares* (OLS), its base form is shown in Equation (52):

$$y = \beta_0 + \beta_1 x, \tag{52}$$

where β_0 is the intercept point, β_1 is the slope coefficient, and x is the time the sample was taken. Here, x can be assumed known, with only the gas values, y, having *Uncertainty*. Assuming they are unbiased, the expected estimates of the intercept, $\bar{\beta}_0$, and slope, $\bar{\beta}_1$, can be obtained using Equations (53) and (54), respectively.

$$\bar{\beta}_0 = \bar{y} - \bar{\beta}_1 \times \bar{x}, \tag{53}$$

$$\bar{\beta}_1 = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{54}$$

where $\bar{x}$ and $\bar{y}$ are the mean values for the samples' time and gas values, respectively. The objective function for fitting the line is to minimise the *Sum of Squared Errors* (SSE) which represents the difference between the observed samples and the estimated trend line, shown in Equation (55). The *Residual Standard Error* (RSE), $\sigma_{\tilde{y}}$, is then remaining average error of the residuals about the trend line, shown in Equation (56).

$$SSE = \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2, \tag{55}$$

$$\sigma_{\tilde{y}} = \sqrt{SSE/(n-2)}, \tag{56}$$

where $\tilde{y}$ is predicted value. Increased RSE can be related with reduced goodness-of-fit via the closely related metric of the coefficient of determination, R^2, shown in Equation (57). The denominator represents the total variability about the mean value, also known as *Sum of Squared Total* (SST), as opposed to being about the trend line as for SSE. This means that a low value of R^2 would indicate a poor characterisation of the gassing rate. Although reduced goodness-of-fit could be related to increased *Uncertainty* of its 'correct' characterisation of gassing rate, given it is the sole metric being used, it is not necessarily implying that trend line is wrong but that it is insufficient to explain the observed variability.

$$R^2 = 1 - \left(\frac{SSE}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}\right), \tag{57}$$

$$\sigma_{\beta_j} = \sqrt{\sigma_{\tilde{y}}{}^2 \times (X^\intercal X)^{-1}_{jj}}\,. \tag{58}$$

The confidence intervals of the coefficients are based on their estimated standard errors, σ_{β_j}, as shown in Equation (58), where j relates to the coefficient index, in this

case being either one or two representing the intercept and slope, respectively. Therefore, with a lot of samples, the confidence in the coefficients can be much greater than the confidence in the predictions. Consider **Gas 3** in Fig. 4-7; it has a $\sigma_{\tilde{y}}$ = 42.05 ppm whilst having a σ_{β_0} = 3.84 ppm and σ_{β_1} = 0.01 ppm/sample. This can be interpreted as there being a large degree of *Uncertainty* in the prediction interval due to the variability of the data whilst the large number of samples indicate that there is little *Uncertainty* regarding that the calculated regression line is the best to fit the data. Note that Fig. 4-7's legend is expressing the 95% confidence interval of the parameters, approximated via $k = 2$ assuming a $\mathcal{N}$ distribution. It would be more rigorous to assume a Student's t-distribution with appropriate degrees of freedom, especially when considering fewer samples. The *Uncertainty* related to goodness-of-fit is arguably irrelevant with regards to the *Uncertainty* of the average gassing rate specifically.

However, even though the *Uncertainty* of the slope coefficient can be propagated directly as a distribution, quantifying it remains challenging. A key aspect to consider is to what extent is the sampled data repeatable in the hypothetical that they could have been sampled again. This is a complex topic as it depends on the assumptions. For example, it may be considered that the ambient temperature is 'random'. In this context, one may wish to include the *Uncertainty* introduced by the added variability caused by the ambient temperature. In contrast, it may be considered that the TX loading is not 'random' and that its effects should be included to the gassing rate but not as a source of *Uncertainty*. A more rigorous approach would be to attempt to keep as many factors as similar as possible, and record those that may change. Then, other methods such as a random-effects model [98], or a multivariate model can be used to attempt to account for these factors.

Note that unlike the delta-based approach, the linear regression-based approach required no input of *Measurement Uncertainty* and instead infers *Uncertainty* on the assumption of normally distributed residuals about a fixed-effect model. However, as this assumption is increasingly violated, the validity of the estimated *Uncertainty* wanes. For example, consider Fig. 4-8. The three lines in red, purple, and blue, represent the presumed *Measurement Uncertainty* of ±15%, the estimated *Measurement Uncertainty* from Fig. 4-4, and the estimated *Measurement Uncertainty* in accordance with the previously outlined equations. **Gas 1** and **Gas 2** both had

residuals resembling a $\mathcal{N}$ distribution and accordingly, both the estimated *Uncertainties* via the linear regression and via Fig. 4-4 are similar, as shown by the blue and purple lines. Furthermore, over time, the much higher ±15% *Uncertainty*, shown as the red line, eventually converges. However, as previously discussed **Gas 3** has significant variability that leads to the *Uncertainty* estimated via the linear regression to be much higher than that of the manually estimated approach. Here, it exceeded the ±15% *Uncertainty* and did not converge.

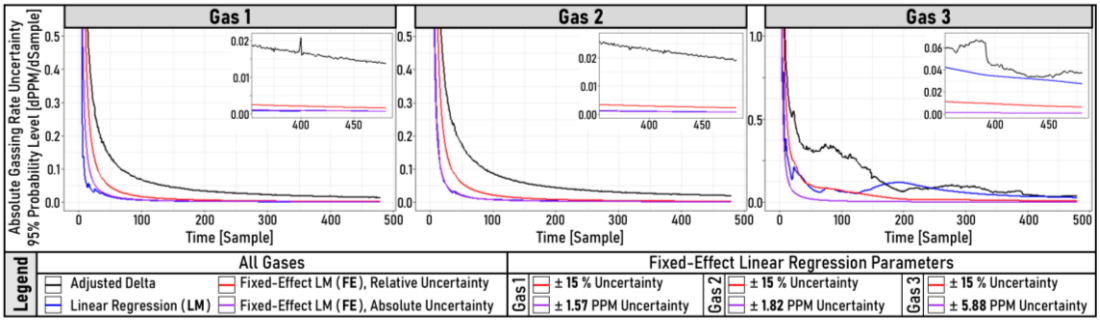


Fig. 4-8: Comparing Uncertainty in Gassing Rate between Delta-Based and Linear Regression-Based

To input the *Uncertainty* explicitly, a *Fixed-Effect* (FE) model can be utilised, where the *SSE* is modified to be the sum of the desired variances of each sample, as was done for Fig. 4-8. This means that by default, there would be no *Uncertainty* unless some variance is input. Here, the *Measurement Uncertainty* is used. Additionally, due to the partitioning properties of summed squares, one can conclude any residual *Uncertainty* can likely be attributed to other sources. Though this is only an estimation as it assumes the observed samples were all the mean values of their true distributions. Additionally, it does not wholly address the previous topic regarding the decision of what, and how much, *Uncertainty* to include. That more complex topic is considered out of scope and instead only the *Measurement Uncertainty* is considered.

Weighting Linear Regression

A topic not yet discussed is whether the linear regression should be weighted. It is relatively common to weight samples in a FE model based on the inverse of their variance based on the intuition that there is a greater degree of *Uncertainty* regarding their true value [98, p. 4], [99, p. 306]. If a fixed *Relative Uncertainty* is assumed as is for the *IEC Specification*, then this would theoretically apply and could account for heteroskedasticity of the *Uncertainty* of each sample. A fixed *Absolute Uncertainty* in contrast, would result in samples being weighted equally and thus have no effect. To

explore this impact, Fig. 4-9 compares weighted and unweighted linear regression models that can be compared to Fig. 4-7. The top row shows the final estimated regression lines using weighted and unweighted in cyan and in blue, respectively. The bottom row shows the estimated gassing rates comparing the two models. **Gas 1** and **Gas 2** seem very similar across both models whereas **Gas 3** is significantly impacted by the choice, with the weighted model seeming to have less variability over time but to settle at a higher estimated gassing rate than the unweighted model.

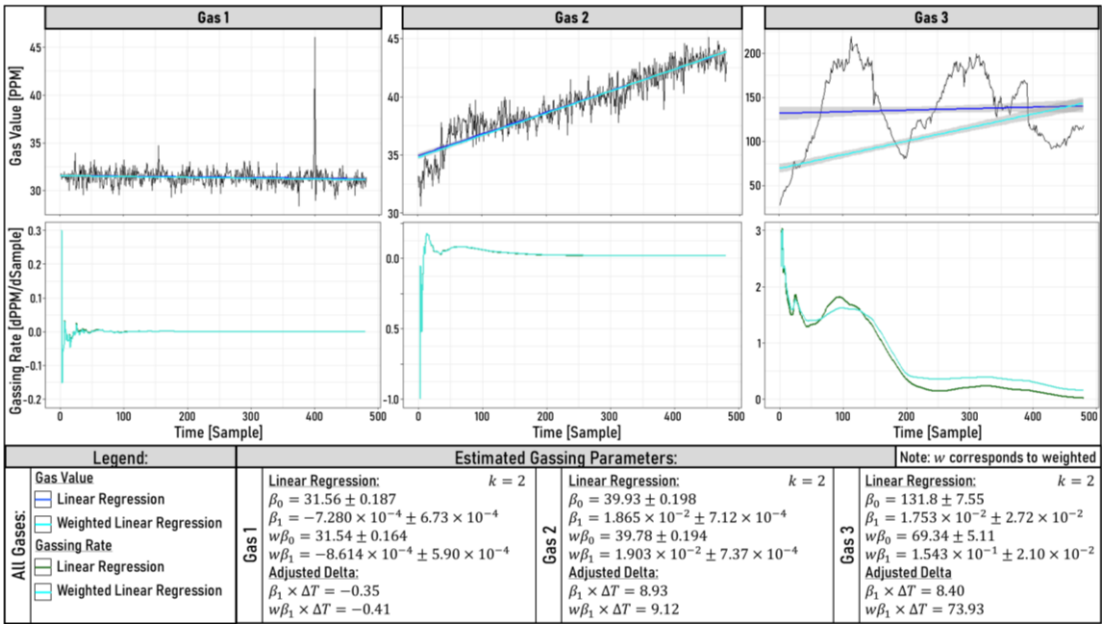


Fig. 4-9: Comparing Weighted and Unweighted Linear Regression

However, this analysis does not determine if weighting improves model accuracy. **Gases 1–3** were records of observed data captured via OLDGA and do not seem to exhibit *Relative Uncertainty*. Therefore, the effect of weighting is investigated on a variant case study. **Gases 4–6** were generated derived from **Gases 1–3**, respectively. They have had a centred 48-sample median filter applied with synthetic *Measurement Uncertainty* superimposed. Three variants are used; the first is without any added *Measurement Uncertainty*, the second has ±15% *Relative Uncertainty*, and the third a fixed *Absolute Uncertainty*. The *Relative Uncertainty* was calculated based on observed values after applying *Measurement Uncertainty* to simulate a practical case. The values used for the *Absolute Uncertainty* are those previously estimated in Fig. 4-4. Using these variants, weighted and unweighted linear models were used to estimate the trend line, assuming inverse variance where applicable. The ideal model would be expected to match the pseudo ground truth of having no *Measurement Uncertainty*. Fig. 4-10

shows the results, with the intercepts and gassing rates shown in purple and black, and the targets as grey dashed lines, respectively.

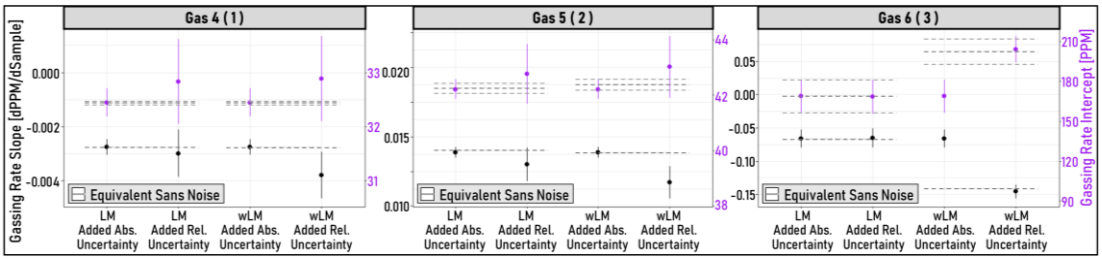


Fig. 4-10: Comparing Uncertainty in Gassing Rate between Weighted and Unweighted Linear Regression

Table 4-1: Studentised Breusch-Pagan Test for Heteroskedasticity of Linear Regressions

Linear Model Type	Studentised Breusch-Pagan Test for Heteroskedasticity [p-value of null hypothesis homoscedasticity]		
	Gas 4 (Gas 1)	**Gas 5 (Gas 2)**	**Gas 6 (Gas 3)**
Unweighted	3.271e-01	4.916e-01	2.200e-16
Weighted	5.087e-06	6.273e-05	4.034e-01

There are two main aspects to consider: to what extent does weighting the model impact the outputs, and then to what extent does the true nature of the *Measurement Uncertainty* influence the outputs. In the three considered examples, weighting the samples had little effect on the data with superimposed *Absolute Uncertainty*. However, it appears to deflate the slope coefficient more noticeably when the data was superimposed with *Relative Uncertainty*. Some of this is explainable by the fact the slope coefficient also reduced from the case with no *Measurement Uncertainty* applied. The weighting seems to then exacerbate this change. Weighting has no effect for *Absolute Uncertainty* but has an unpredictable effect on performance in the case of *Relative Uncertainty*. Both **Gas 4** (**Gas 1**) and **Gas 5** (**Gas 2**) were hampered by the weighting in these cases whereas **Gas 6** (**Gas 3**) saw its estimate greatly improved. Another aspect to consider is whether weighting decreased the heteroskedasticity in the residuals as intended. Table 4-1 tabulates the results of a Studentised *Breusch-Pagan* test [100] for both the weighted and unweighted models, which is a test for heteroscedasticity. This was implemented using the "bptest" function in the "lmtest" package in R [101]. The results show that weighting samples can both increase and decrease heteroskedasticity. It should therefore not be assumed that weighting will improve performance, and in this thesis, it is thus not used.

Impact of Samples Placement on Uncertainty

It has thus far been assumed the model was privy to all samples, at least within the consecutive window with which it was basing its trend. However, this is only representative of OLDGA: lab-based analysis will have far fewer samples. Drawing from [1], one might assume indicatively between 3–6 samples over the relevant duration, here being assumed as 2 years. In this context, the estimates may differ drastically depending on the number of samples and the time that they were taken. It may be presumed that having equidistant sampling is optimal given the increased coverage of the period. However, for linear regression, this intuition does not apply to the slope coefficient specifically.

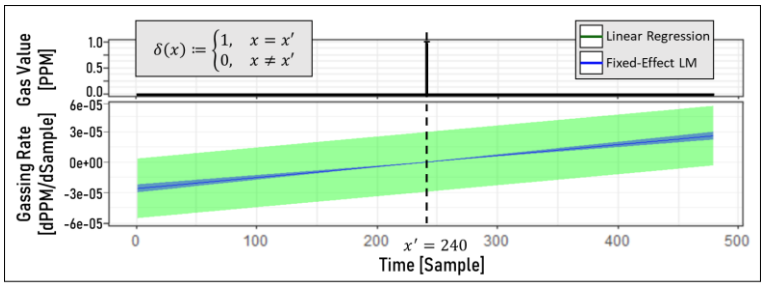


Fig. 4-11: Comparing Delta-Based Gassing Rate to Linear Regression using pulse function

For example, consider Fig. 4-11 showing a Kronecker delta function [102], $\delta_{xx'}$, that pulses at point x' from zero to one. The *Uncertainty* using the linear model and FE model are shown in green and blue based on ±15% *Uncertainty* at a 95% probability level, respectively. If calculating the *Uncertainty* of the gassing rate by explicitly providing *Measurement Uncertainty*, the time at which the pulse is applied impacts the slope coefficient's *Uncertainty* such that the nearer it is this to the mean value of samples times, the less influential it is. This is rather counter-intuitive as one would at first presume the most natural approach to adding samples is placing them at the mid-points of existing intervals. This may be due to our implicit assumption of some (positive) autocorrelation between samples taken at a short interval.

Impact of Number of Samples on Uncertainty

The final aspect considered is the impact of number of samples used: thus far, many of the examples have been using 480 samples. Referring once again to [1], it estimates gassing rates for its Table 4 using 3–6 samples. Similarly, the previously discussed [66, p. 23] suggests even 3–4 samples can make the accuracy of the gassing rate "generally much better". In this context, the estimates may differ drastically depending on the

number of samples and the time that they were taken. However, this is a very open-ended topic dependent on the assumptions being made, making it difficult to fully address. Given the first and last sample, the impact of additional intermediate samples are examined here.

Therefore, keeping the first and last sample fixed, an additional 4 intermediate samples were randomly picked using a uniform distribution from **Gases 1–3**. This was repeated 10^4 times to form an empirical distribution using between 3–6 samples to compare their estimated gassing rates. The same sample indices were used across the gases, and the outputs are shown in Fig. 4-12, where each row represents one of said three gases. Only the estimated gassing rate is shown, as the estimated intercept is irrelevant to the IEEE methodology [1]. The left column shows the empirical distribution of the mean value, and the right column shows it for the *Uncertainty*. Although it can be misleading to decouple the *Uncertainty* from the estimated value, it is done here simply to highlight general trends. Line colours indicate the number of samples used to generate the distributions. The marks, either plusses or circles, represent the method used to estimate the parameters. The first uses simple linear regression and the second is the FE model: only estimates of *Uncertainty* will be affected this choice. This can be seen in the left column, where the distributions for either method align. Lastly, the two vertical lines in purple and green represent the output when all 480 samples are used for the two methods, respectively.

If one considers the purple and green lines in Fig. 4-12 where all samples were present as the pseudo ground truths, then the hope would be increasing the number of available samples leads to distributions rapidly converging to these values. Although, it is slightly less clear the desired behaviour for the *Uncertainties*. Either one could argue for an estimation matching the pseudo ground truths as closely as possibly, or that it should be estimated at a higher value to reflect the fewer available samples.

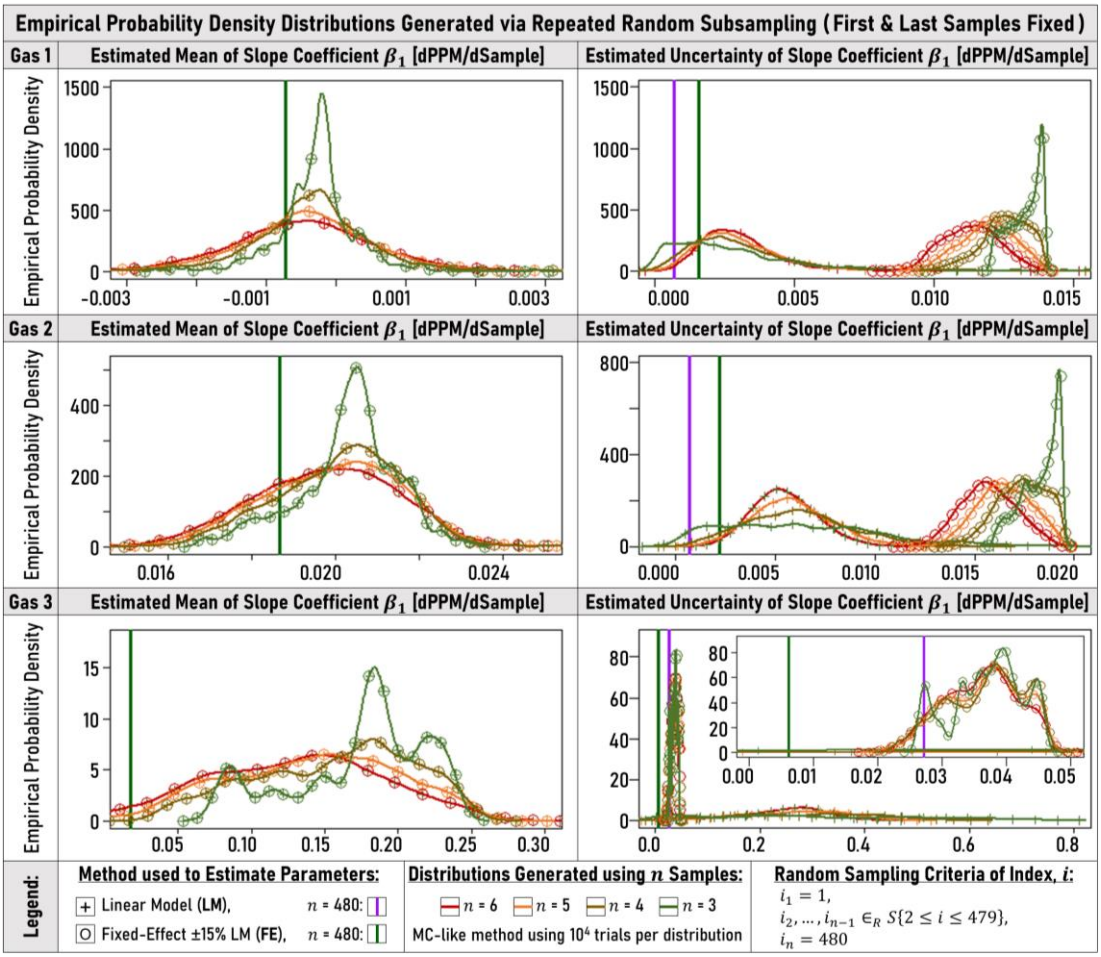


Fig. 4-12: Comparing Estimated Parameters Based on Number of Samples

When considering the estimates of the mean on the left of Fig. 4-12, **Gas 1** clearly trends towards debiasing as sample count increases. The case for **Gas 2** is less clear as it has not yet converged by the time 6 samples are used. **Gas 3** appears more clearly to **not** be trending towards debiasing to a significant extent by the time 6 samples are used. The same patterns are followed for the estimates of *Uncertainty* when using the simple linear model, shown on the right of Fig. 4-12 using plus markers and the green line as the pseudo ground truth. However, if using an FE model, they all appear to behave relatively consistently regardless the gas; estimating at a much higher value than the case of having all samples available, and slowly debiasing as the number of samples increase. This is shown using the circle markers and the purple line as the pseudo ground truth. As the FE model's estimates of *Uncertainty* is heavily impacted by the number of samples, its overall behaviour is as expected. Care should be taken when comparing the estimates of *Uncertainty* against the estimates of the mean of the gassing rate as the FE model is explicitly ignoring variability. This means that where

variability is high, such as **Gas 3**, it is expected for it to underestimate overall *Uncertainty*. However, for **Gas 1** and **Gas 2** where there is low variability, ±15% *Uncertainty* could be considered an overestimation.

Across all the metrics, increasing the number of samples from 3 to even just 4 dramatically altered the distributions towards unbiased distributions, and by 6 samples they all seemed relatively unimodal distributions. Although, the overall variability or spread of the distributions increased in the debiasing process. The extent to which the gassing rate, as described by solely the first and last sample, is in alignment with the case of having all samples available is the primary driver to the difference in these predictions. The rate of convergence is then also driven by the overall variability of the gas levels, where a greater variability slows convergence. Fig. 4-13 considers **Gas 3** in a similar approach except that **all** samples are randomly selected as opposed to having the first and last samples fixed. The effect is that there is no longer a disproportionate skew towards said two samples, and the rate of convergence seems more in line with **Gas 1** for example. It should be noted that Fig. 4-13 has little meaningful relevance to a practical scenario. Perhaps a more ideal setup, though not explored here, would be comparing the outputs of randomly selected samples against having all intermediate samples spanning the range of said randomly selected samples. However, as this range varies each random sampling, so would the relative proportion of "coverage" a sample represents, requiring some form of normalisation.

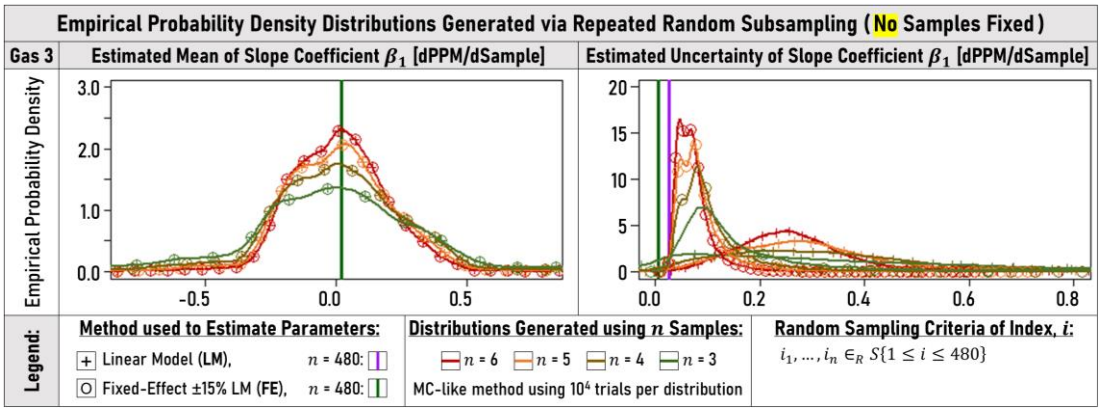


Fig. 4-13: Comparing Estimated Parameters Based on Randomly Distributed Number of Samples

Increasing the intermediate samples can improve estimation, quickly converging on the final estimate when there is low variability, as seen with **Gas 1** and **Gas 2**. However, 6 samples cannot be assumed generically sufficient to converge, as seen with **Gas 3**.

4.1.5. Findings

Absolute Gas Levels

Although the inclusion of *Uncertainty* can influence the confidence given to an output, the additional nuances seem to have diminishing effects. It is mainly only relevant for C_2H_2 and H_2 to consider the increased *Uncertainties* due to proximity to LoD, and then only relevant for C_2H_2 to consider the inclusion of a minimum *Absolute Uncertainty*. H_2 is not only affected to a lesser extent than C_2H_2, but it is already known to be a more unreliable gas for *Diagnostics* due to reasons previously discussed whereas C_2H_2 is considered a critical gas. C_2H_2 presents a well-known challenge in DGA as the expected concentration values and thus the limits set by many approaches are typically $<10 \times S$. At the extremes, where both the LoD and the limit is at 1 ppm, the minimum value needed to be detectable is already at the limit. This results in significant *Uncertainty* when assessing C_2H_2 at these levels.

It is highlighted that [68, Sec. 9] makes use of an *Absolute* minimum for *Repeatability* but then not for *Accuracy*, and that this is difficult to justify. However, as this thesis is using only the *Accuracy* given in [68, Sec. 9], the *Absolute* minimum is not considered. Furthermore, neither [1], [2] make mention of an *Absolute* minimum for *Accuracy*.

Relative Gas Levels

There are many nuanced complications when applying linear regression to DGA samples en masse as each model's validity is dependent on the input data. The blanket application of a linear regression can have unpredictable results if the validity of said model is disregarded. However, utilising an estimate of the average gassing rate over the duration can provide extra consistency useful for comparing metrics. Furthermore, the alternative of using simply the delta does not address these issues. It can be concluded that there is a clear enough distinction between the outputs of estimated gassing rate via linear regression than from a delta-based approach that they should be considered different metrics. Therefore, though the generic advice that a linear regression may be "better" for characterising gas rates is defensible, it should also be caveated with a caution against it being considered a like-for-like replacement to a metric characterising the gas accumulated over the given period.

Often, ±15% *Accuracy* is assumed as the default where no further information is available as per the *IEC Specifications*. However, this assumption of a fixed *Relative Uncertainty* has complicating implications. For example, it would imply a bias if assuming a $\mathcal{N}$ distribution about the observed sample rather than about the unknown 'true' sample value. Furthermore, it may suggest a weighted model more appropriate to account for the heteroskedasticity inherent to *Relative Uncertainty*. The limited samples of real OLDGA data observed seemed to suggest that the *Precision* did not scale linearly with gas level. If it is not a linear relationship, perhaps it may be better represented with either an *Absolute Uncertainty*, or as a combination of the two. It was demonstrated in [33] that the *Relative Accuracy* of OLDGA was worse at low gas levels, but whether results based on two levels should be linearly interpolated is unknown.

Based on a dataset of just three examples, it can already be concluded that the predicted gassing rate and associated predicted *Uncertainty* seem somewhat unreliable if taken at face value. They should rather be seen as indicative metrics used for comparative purposes rather than faithfully representing the 'true' gassing rate and associated *Uncertainty*. This is particularly true where there is variability in the gas levels beyond the linear trend. In these cases, removing the variability would of course help but presumes the *a priori* knowledge of its presence, which can be unrealistic when dealing with as few as 3–6 samples.

4.2. Case Study Analysis of Methodologies

4.2.1. Scope

This Section applies automated implementations of the reviewed DGA methodologies developed for this thesis to real TX DGA data to explore their relative behaviours and identify potential barriers to practical deployment. The two main DGA methodologies considered are IEEE C57.104-2019 [1] and IEC 60599:2022 [2]. Additionally, the NEI method outlined in [1, Annex F] is considered and compared to the LSA method [5].

There are potentially many outputs, comparisons, and avenues of analysis to consider. The intention is not to definitively interpret the DGA data to *Diagnose* the TXs as ground truth is unavailable, but it is similarly difficult to avoid the topic entirely. Therefore, it should be noted interpretations are based on conjecture and/or literature introduced thus far in the thesis. Rather than present every possible output, specific

topics will be highlighted alongside relevant plots demonstrating the point drawing from the case studies. The main topics of interest are:

- difficulties arising automating the methodologies,
- potentially under-specified edge-cases,
- systematic behavioural differences between the methodologies, and
- difficulties in applying methodologies where sampling rate varies, and for OLDGA.

These novel outputs can inform would-be users of a methodology of its respective expected behaviour as compared to the alternatives. Furthermore, they provide demonstrative examples using real TX DGA of potential barriers to deployment.

4.2.2. DGA Interpretation Methodology Implementations

IEEE C57.104-2019 Implementation

A strict automated implementation of IEEE C57.104-2019 is impractical given the nuances and the inherent subjectivity it advocates. As per [1, Sec. 6.1] "DGA interpretation is still more of an art than a science…". Nevertheless, the reality is a primary motivator for *Screening* is the excess of data, suggesting a need for automated procedures. Therefore, a faithful but simplified implementation is attempted, with a detailed explanation of the interpretation decisions.

It is challenging to formalise the derivation of the *DGA Status* levels comprehensively given the amount of discretion [1] advises regarding the classification. There are numerous edge cases where it suggests overriding the *DGA Status* based on various factors. The original work for the automated implementation related to the thesis was published in [75]. Fig. 3-2 is modified from [75, Fig. 1] to include the derivation for the cases where $\mathbb{T}3\text{–}4$ may be incalculable, as based on [1]. This derivation process is repeated for every available gas. The worst case amongst all available gas outputs, **L**, is selected to then determine the overall *DGA Status* level, **Ł**, of the given sample.

As was discussed in Sub-Section 3.1, a complicating factor of the derivation process is the inclusion of the *Confirmation Sample* as this can retroactively impact the validity of prior samples depending on its output. Steps 7 and 8 of [1, Sec. 6] were not implemented here. The first states that if a TX is **Ł3** for a prolonged period without significant gassing, then a lower *DGA Status* level may be considered. The second states that for "extremely high concentrations, deltas, or rates", an expert should be consulted.

Diagnostics is not a focus of this thesis and nor does the methodology outlined in [1] contribute significantly towards advancing this topic. Therefore, for simplicity, only *Duval Triangles 1*, *4*, and *5* were implemented from those explicitly mentioned in [1]. *Duval Pentagon 1* and the *Rogers Ratio* method were omitted as they were considered to overlap in scope with the *Duval Triangles* and the *IEC Ratio* method, respectively. The *IEC Ratio* though not mentioned within [1], will be implemented as it is required for the IEC 60599:2022 methodology.

The initial automated implementation process, done in R, is shown in Fig. 4-14 and represents the "base" methodology. To calculate the delta in gas levels, at least two valid samples are required. To calculate the gassing rate, at least three valid samples are required, in addition to the T4 sample selection criteria. The implementation assumes resamples occurred wherever the guidance suggested it. The data is input as .csv files and the outputs are similarly saved. Visualisations were done at runtime in R. The packages "ggplot2" [103] and "ternary" [104] were used to aid with this. A validation of the implementation using the case study examples provided in the Annex of [1] are provided in Appendix 4 of [65].

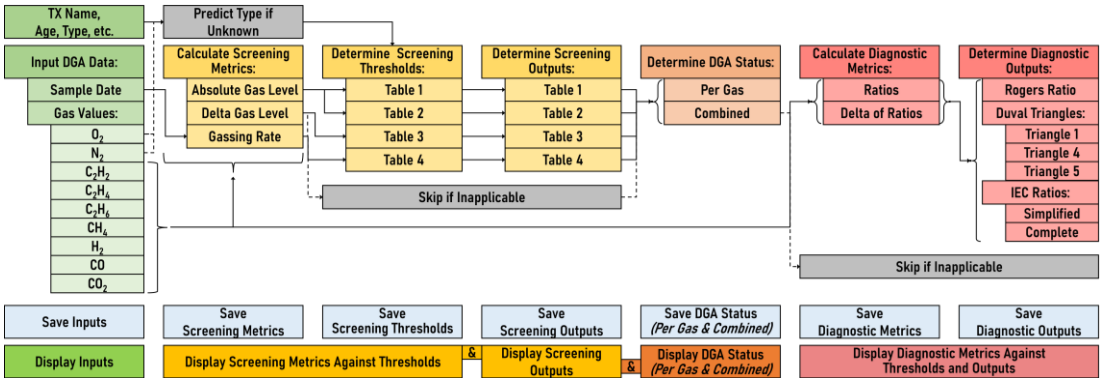


Fig. 4-14: Interpreted Automated Implementation of IEEE C57.104-2019's Methodology

IEC 60599:2022 Implementation

The automated implementation, done in R, follows the logic shown in Fig. 3-4 whilst following the same general approach as for the IEEE methodology shown in Fig. 4-14 with the notable exception that the *Diagnostic* process is done to determine the final *Screening* output, unless it is classed as in a *Typical* state. Like the IEEE implementation, the data is input and output as ".csv" files and the packages "ggplot2" [103] and "ternary" [104] were in R used to aid with visualization of plots.

If the DGA values are below ten times the analytical limit, this implementation will still perform the analysis but with a flag attached to the output, warning the engineer of the increased *Uncertainty*. In cases where the DGA value is below the analytical limit, said limit will be assumed to allow for the calculation of the ratios, again with the addition of a flag warning of the increased *Uncertainty*. Although [2, Sec. 9] states that inconsistent DGA values must be rejected or corrected, this implementation assumes the engineer addressed this prior to running to analysis.

It is unclear how to interpret when to perform ratio checks, [2, Sec. 6.1] states:
"Gas ratios are significant and should be calculated only when at least one of the gases is at a concentration and a rate of gas increase above typical values (see Clause 9). Nevertheless, it is recommended to also calculate them in cases where one or more gases show increasing or abnormal concentrations, even if they are lower than typical values. Avoid calculating ratios when the gas concentrations are not high enough to be reasonably accurate according to IEC 60567".

However, [2. App. A.2.3] states "any formation below typical values of gas concentration and rates of gas increase should not be considered as an indication of "fault", but rather as "normal gas formation". Ratios are not significant in such a case". This is echoed in [2, Sec. 5.1]. It is difficult to interpret whether the ratios should be calculated and are significant under these situations.

Therefore, the ratio checks are performed whenever a gas increases and flagged when at least one metric is above the typical values for **both** the gas concentration and gassing rate as initially recommended. It is stated in [2, Sec. 6.1] if gas ratios are different to previous analysis to consider using deltas to calculate the ratio. Therefore, the process is repeated twice, once with absolute values, and once with the delta from the greater of one month prior or the previous sample. It is expected of the engineer to discern which output is more relevant for the given case. The choice of one month is rather arbitrary, but as [3, Sec. F] states, monthly values are "an intuitively good time scale for expressing gas accumulation rates in transformers" and though these ratios are not accumulation rates they are both attempting isolate the delta.

As discussed, not all limits are provided in IEC 60599-2022. Therefore, L_1 / G_1 was used as the *Typical* quantity, and L_2 / G_2 as the *Intermediate 2* quantity. These are tabulated

in Table 3-1 and Table 3-2 from Tables C.7 and C.8 from [62], respectively. Another difference between the methodologies outlined in [1] and [2] is that the latter suggests a distinct method for *Fault Identification* whereas the former refers to existing methods. Therefore, *Fault Identification* is also considered an output here. The *IEC Ratio* method and the previously discussed *Duval Triangles 1*, *4*, and *5* were implemented.

To address the instruction in [2, Sec. 9] to "verify if fault is evolving towards final stage", the results of the ratios are plotted to allow for graphical interpretation of the "trajectory" to determine whether it appears to be heading towards a more severe *Fault*. The process is also repeated twice, once with absolute values, and once with the discussed delta. Inferred from [2, Fig. 1] (Fig. 3-4), the *Alarm* condition can be reached either if a gas concentration **and** rate of gas increase are above *Alarm* values, or if they are above the *Typical* values whilst a *Fault Type* of **D2** is indicated. If they are both above the *Typical* values but below the *Alarm* values without indication of a **D2** *Fault Type*, then an *Alert* condition is outputted.

To address the instruction in [2, Sec. 9] to "determine if paper is involved", two checks are done as per [2, Sec. 5]. The first is if the CO > 1,000 ppm whilst $CO_2/CO < 3$, and the second is if CO_2 > 10,000 ppm whilst $CO_2/CO > 10$. The first constitutes an indication of paper involvement with possible carbonisation, and the second an indication of mild overheating of paper or oil oxidations. Although, these are only relevant in cases where H_2 or other hydrocarbons are present to corroborate these checks. These checks are outputted in the form of flags and do not override other *Diagnostic* outputs. The *Duval Triangles* are also used to indicate paper involvement.

Lastly, for addressing the instruction in [2, Sec. 9] to "take proper action according to best engineering judgment and/or with help of Figure 1". The output states whether, and if so which, *Typical* and/or *Alarm* values were exceeded. The consequence is left to the engineer to discern. The *Diagnostic* plots are also provided to help with the interpretation of the outputs. Guidance on how the information should be output is provided in [2, Sec. 10], but many of the details, such as the sampling location was not known for the subsequent case studies, but it is assumed they can be trivially included in the outputs alongside the implementation in future work.

Normalised Energy Intensity Implementation

It is stated in [1, Annex F] that action may be determined based on outputs of *Fault Type Identification* and the NEI. However, it is not entirely clear how to implement this. It is in this thesis interpreted that this is not intended for *Screening*, so once **Ł2–3** is reached, then the NEI values are tracked, mirroring the guidance related to *Diagnosis*. Also stated in [1, Annex F] is that "separate attention should be paid" to CO, CO_2, and C_2H_2. However, it is again not clear how to implement this. As per [1, Annex F], "if NEI_{paper} is increasing, especially if the CO_2/CO ratio is also significantly decreasing, there may be a fault affecting insulating paper", but there is no subsequent definition as to what constitutes a "significant" decrease in the CO_2/CO ratio. Similarly, there is no guidance regarding how to make use of the tracked C_2H_2. The only note regarding limits was the following, as per [1, Annex F]:
"Experience with this NEI-based method at a large US electric utility suggests that an NEI_{oil} increment of 0.5 or an NEI_{3oil} increment of 0.3 over any time interval should raise concern for the transformer's condition, and larger increments warrant correspondingly more concern".

Automating this is challenging due to the lack of nuance. Once a limit is reached, if it is decided that the current state is non-problematic, it is not clear how to proceed: should the limit now be 'reset' taking here as the new norm, or should the flag remain indefinitely. In both cases, resetting the limit to be relative to the current value seems ill-advised; rather, engineering judgement should be applied to select an appropriate *Baseline* manually. A typical approach to avoid these scenarios is via the use of a rolling-window which inherently accommodates gradual gas accumulation.

A single combined implementation was chosen to address the three potential NEI methodology sources: [1, Annex F], [3], and [4]. This was challenging as the suggested values seem irreconcilable in places. For example, [1, Annex F] has no guidance for the absolute value of the NEI whereas [4] has no guidance for the rate of change, only absolute value. Although [3] did have guidance for the rate of change for NEI_{3oil} in the form of the 90^{th} percentile of the per month change for three populations: 2.0, 8.2, and 4.7. However, when this is contrasted to [1, Annex F]'s lifetime limit of 0.3, there seems a large discrepancy, especially considering [3] is based on just a one-month interval. Furthermore, the published summary statistics of the source datasets used in [3] and

[4] seem very different. For example, comparing the 90th percentile values for CH_4, C_2H_6, and C_2H_4.

There is a clear logic to tracking both absolute values and rate of change as was suggested by [3]. Similarly, they had claimed to observe a large difference in values by O_2 level in [4] which can be partly corroborated by the [1] stratifying limits based on O_2 level. It is therefore not clear why their latest rendition published in [1, Annex F] then ignored both rate of change and differentiating by O_2 level. Given that the three sources were apparently all led by the same authors, this thesis assumes that over time the dataset was refined, and that the IEEE documentation had it simplified.

The initial automated implementation process, done in R, is shown in Fig. 4-15 and represents the "base" methodology. It is therefore heavily skewed towards the IEEE documentation. The guidance provided by [1, Annex F] was rather sparse and thus required some assumptions to create Fig. 4-15. It was interpreted that NEI_{paper} is primarily to inform *Diagnostics*, although it can influence *Fault Severity* assessment as paper involvement is considered a higher priority. Similarly, it was interpreted that the lifetime-increment limits for NEI_{oil} and NEI_{3oil} are a *Screening* metric rather than a *Fault Severity* one, even though the absolute NEI values are used for *Fault Severity*. Ultimately, the taxonomy is not too impactful on the implementation and its outputs.

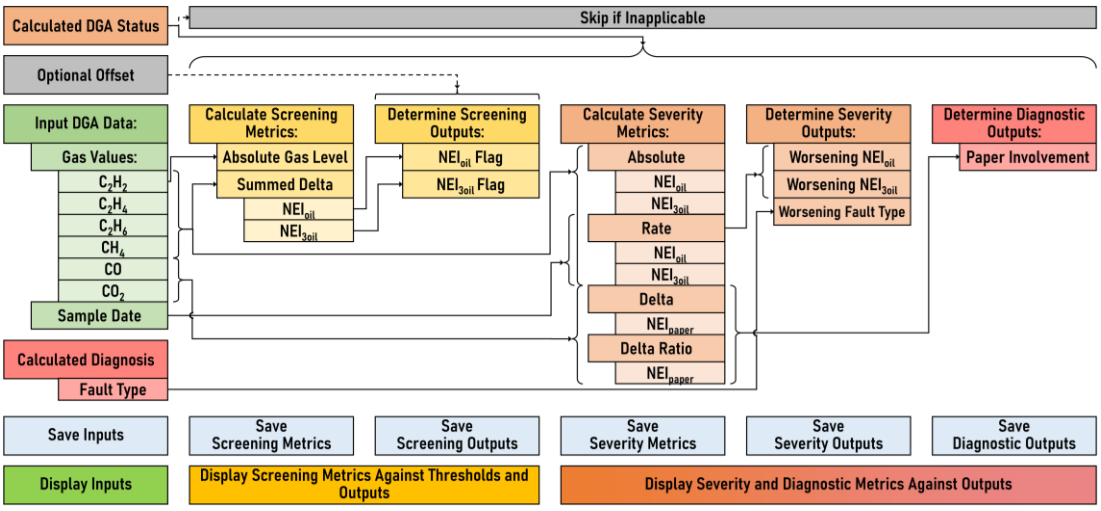


Fig. 4-15: Interpreted Automated Implementation of IEEE C57.104-2019's NEI Methodology

Both NEI_{oil} and NEI_{3oil} will be calculated, and it is left to the engineer to discern when to use the latter by identifying when "significant ethane stray gassing is suspected" [1, Annex F]. Similarly, C_2H_2 values are shown for every sample, and it is for the engineer

to discern the relevance. Any derivative calculations will be based on NEI_{oil} only. As previously mentioned, this thesis assumes that the NEI methodology is applicable according to the same criteria as the *Diagnostics*, i.e., only once **Ł2–3** is reached are the NEI values tracked. The cumulative delta of the NEI_{oil} and NEI_{3oil} is tracked and when exceeding the limits of 0.5 and 0.3, respectively, a warning is raised. If the delta in NEI values per unit time increases whilst in these DGA Status levels, then a worsening situation is assumed. Similarly, if there is a transition from *Fault Types* of **PD**/**T1**/**T2**, to *Fault Types* of **T3**/**DT**/**D1**/**D2**, then a worsening situation is assumed. Additionally, if there is both a positive delta in the NEI_{paper} and a negative delta in the CO_2/CO ratio, then it is flagged that the *Fault* may include insulating paper.

Lapworth Scoring Algorithm Implementation

The implementation used the confidential algorithm. No *Diagnostic* guidance is provided, and the same metric is used for both *Screening* and *Fault Severity* assessment. The metric is converted into an alarm level or flag using the limits.

4.2.3. Case Study Data

Two datasets are considered. The first is three TXs of similar type and loading that were manually sampled over approximately 3–3.5 years, named **TX-A**, **TX-B**, and **TX-C**, respectively. The second is named **TX-D**, which had OLDGA from which a 2-year period is used. There is further information on **TXs A–C** in [105]. "Q" and "Y" is used for referring to the quarter and year shown in the plots, respectively. These case studies were selected from a slightly larger pool of 14 TXs and represent the most insightful examples.

Case Study 1: Overview

Fig. 4-16 shows an overview of the gas levels in all the TXs. The top row shows that absolute gas values for the hydrocarbons with their relative proportions plotted on the second row. The third row shows the absolute gases for the carbon oxides with their relative proportions plotted on the fourth row. A cursory inspection of the carbon oxides shows **TXs A–C** accumulating gases with **TX-C** ending with substantially lower overall gas levels as compared to **TXs A–B**. After approximately Q2Y3, there is a noticeable shift in behaviour with increased gassing rates. **TX-B** has data available for a slightly longer duration than the other two and shows the eventual degassing that

occurred. **TXs A–C** had an initial period of highly volatile results for unknown reasons before stabilising as Y1 begins.

Case Study 2: Overview

Fig. 4-16 shows the overview for **TX-D** in the right column. It indicates generally that the overall hydrocarbon gas levels are quite a lot higher than the other TXs, though with less H_2. The overall gas levels seem more stable, with a gassing rate more like **TX-C** than **TXs A–B**. This is especially noticeable with the carbon oxides. The gas compositions also seem relatively stable barring two noticeably discontinuities: one near Q3Y0, and one near Q1Y2. The first was a large, almost immediate, decrease in H_2 levels. The second was an approximate two-week gap in sampling, followed by a two-day period of elevated H_2 before a sharp drop nearer to previous levels. The cause of these is unknown, however given that the other gases remained relatively consistent, and that they were repeated across multiple samples, it is assumed valid data. Therefore, no remedial action was attempted on the data for this aspect. However, Fig. 4-16 also highlights a likely issue with data quality for **TX-D** which used OLDGA. Samples that had the sum of hydrocarbon gases equal to zero are assumed to incorrect and are removed. This happens primarily around Q2Y0. **TX-D** is used only to explore the impact of varying the sampling rate and not for assessing the outputs directly.

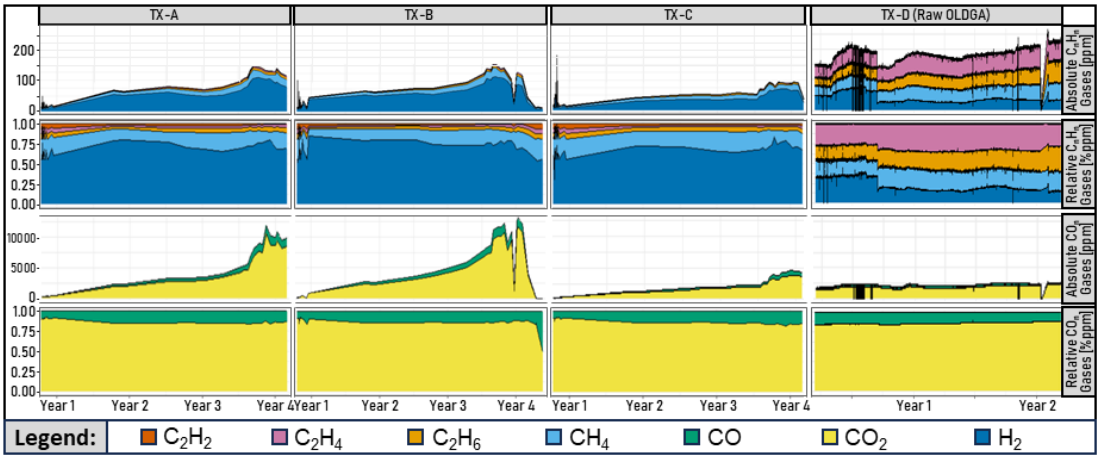


Fig. 4-16: Gas Levels in Case Studies

4.2.4. Results and Findings

NEI_{oil} Compared to LSA

As mentioned in [4, Sec. B], H_2 often seems to dominate stacked plots of raw gas values and can obfuscate other trends, as seen for **TXs A-C** in Fig. 4-16. Fig. 4-17 demonstrates

this by using **TX-A** to compare the stacked plot of hydrocarbons with and without H_2 in the left and middle plot, respectively. The left plot appeared relatively stable between Y2–3 whereas the middle plot shows the hydrocarbons were steadily increasing. The plot on the right of Fig. 4-17 shows the hydrocarbons scaled in accordance with NEI (Equations (2) and (3)) and demonstrates the added emphasis placed on gases such as C_2H_2 which are often considered more important gases. The overall trend of the NEI_{oil} metric seems similar to using raw gas values and is quite intuitive to interpret. This is shown in the bottom row of plots in Fig. 4-18. If considering the 0.5 limit for NEI_{oil} mentioned in [1, p. 82] as an absolute value, then only **TX-D** exceeded it. However, if tracking relative to the initial sample serving as a *Baseline*, then **TX-D** would also not flag. **TXs A–C** were not close throughout the duration, indicating that either none of the TXs were problematic or that the limit is for more severe gassing.

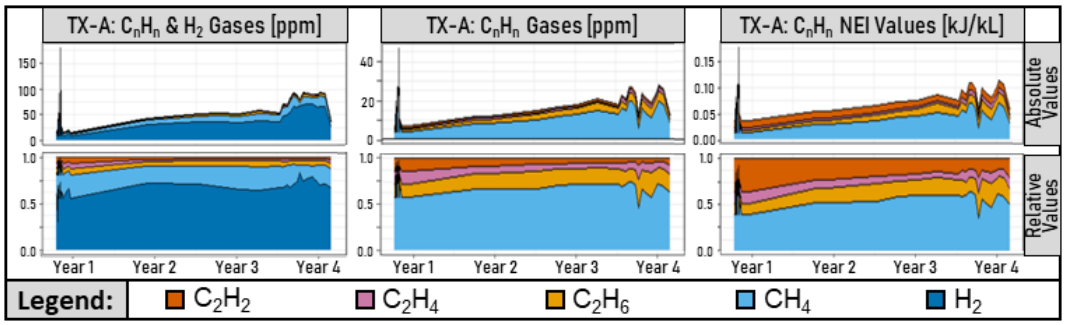


Fig. 4-17: Comparison of Stacked Plots: All Gases, Hydrocarbons, and NEI-Scaled Hydrocarbons

In contrast, the LSA metric shown in the top row of Fig. 4-18 flags near the end for **TXs A–C** where elevated gas levels were observed, as to be expected. It also differentiated between the **TXs A–B** from **TX-C** that had lower gas levels. However, all TXs initially flagged sporadically at the maximum level, with some persisting at a medium level beyond the initial volatility. It is not easy to interpret why this is given the nature of the non-linear function. **TX-D** remained at a medium level flag consistently throughout the duration despite its steadily accumulating gas levels, in part due to its relatively stable composition. The increased sampling interval for **TX-D** will have no effect on both the LSA and NEI metrics and primary flags as they are a function of current absolute gas values. The second row in Fig. 4-18 plots the raw LSA metric, and it can be noted that the trend is somewhat counter-intuitive as it begins very high then lowers overtime, before rising again. The sagging in the middle of **TXs A–C** is due to the disproportionate accumulation of CH_4 during the interim. **TX-D** instead has an unexplained step-change

in H_2 levels. At this point the NEI_{oil} and LSA metrics deviate from one another as only the latter uses H_2. This drop is also the cause of the LSA flag not rising over time.

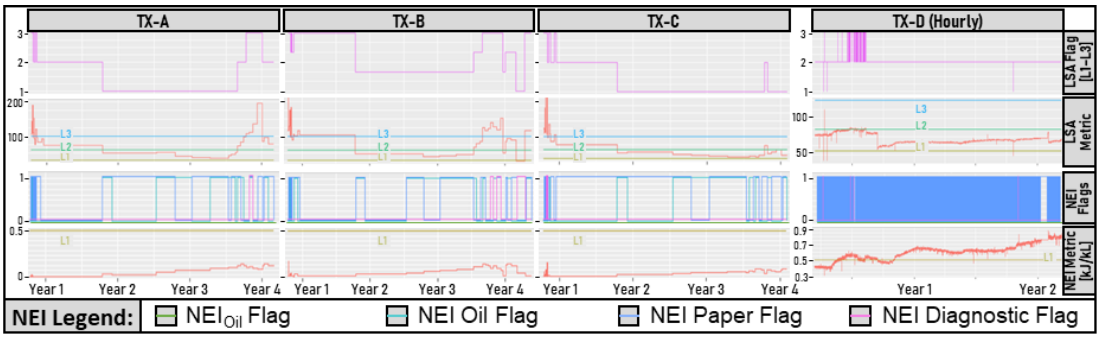


Fig. 4-18: Case Study LSA and NEI Outputs

Another side-effect of the LSA metric being focussed on composition as opposed to purely gas values is that at the end of the data for **TX-B**, it is flagging maximally even though the overall gas levels have dropped substantially post degassing. This highlights the need to note relevant context when assessing this metric and indicates complimentary metrics, such the NEI_{oil}, can be very useful in providing such context. In this example, it can be seen clearly in the NEI_{oil} that the gas levels have dropped, and thus that the LSA flag can perhaps be ignored. As a corollary it can also be concluded that the NEI_{oil} and LSA metrics are substantially different and not interchangeable due to the latter's emphasis on gas composition and inclusion of H_2. An alternative approach to avoid false flagging may be to employ a *Screening* step prior to inspecting the LSA flags and treating it as a *Fault Severity* metric.

Other checks using the NEI were mentioned in [1, Annex F], these are here termed "auxiliary flags", named as *Oil Flag*, *Paper Flag*, and *Diagnostic Flag*, respectively. The *Oil Flag* was based on whether the NEI_{oil} metric was accumulating at an increasing rate and signifies a potentially worsening situation. The *Paper Flag* is based on whether the NEI_{paper} was increasing whilst the CO_2/CO ratio was decreasing, if so, it signifies potential paper involvement. The *Diagnostic Flag* was based on whether *Duval Triangle 1* indicated a progression from *Fault Type*(s): **PD** / **T1** / **T2** to *Fault Type*(s): **T3** / **D1** / **D2**. These auxiliary flags appear in the third row of Fig. 4-18 but seem challenging to interpret. It can be reasonably argued that they are inhibited by the chosen visualisation. Nevertheless, they appear very sporadic and not well integrated into a decision-making programme, suggesting that the implementation of NEI proposed by

[1, Annex F] lack sufficient detail. These flags are sensitive to sampling rates. **TX-D** shows oscillatory flagging likely due to noise, complicating meaningful interpretation.

IEEE Compared to IEC: Primary Outputs

Although [2]'s TX condition outputs and [1]'s *DGA Status* levels share similarities, they lack alignment within this limited case study, as shown in Fig. 4-19. The methodologies' outputs were previously discussed in Sub-Section 3.1. It appears that the [1]'s *DGA Status* is much more prone to outputting its maximal level, though this is of course dependent on the data being used. One possible reason may be that the chosen interpretation for [2] requires both metrics to be above a given limit whereas [1] does not. The bottom row of Fig. 4-19 shows the outputs had an alternative implementation for [2] been explored, where if either metric exceeded a limit, the output level would have been escalated. This approach produces outputs more similar to [1], but with some marked differences such as less frequent maximal-state outputs. Further examination of this alternative implementation for [2] is not considered.

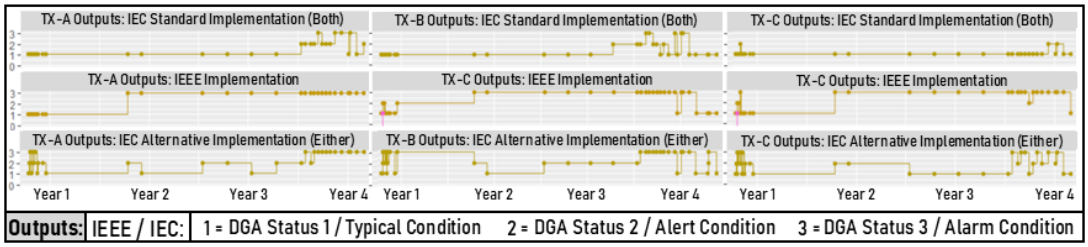


Fig. 4-19: Case Study IEEE and IEC Screening Summary Outputs

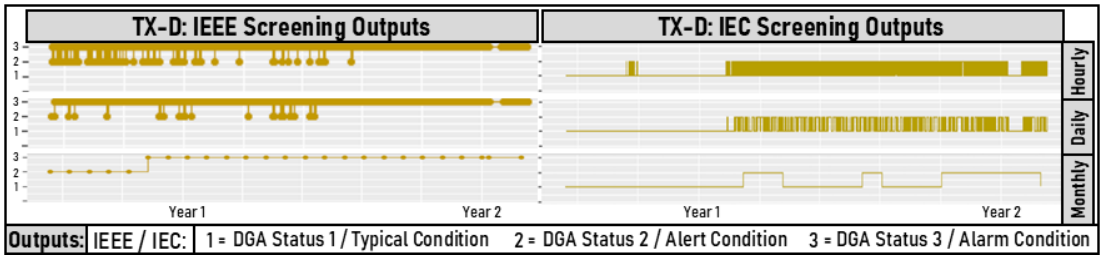


Fig. 4-20: Case Study (TX-D) IEEE and IEC Screening Summary Outputs Relative to Sampling Rate

The *DGA Status* often reached its maximum, limiting scope for escalation or granular comparison. For example, it is unclear which TX is overall in worst state when looking at the *DGA Statuses* in Fig. 4-19. In contrast, [2]'s outputs would indicate that **TX-C** is in a generally less severe state which aligns with both the LSA and NEI metrics shown in Fig. 4-18. Additionally, there is sufficient granularity to capture the potentially worsening situation that began near Q2Y3 whereas the *DGA Status* was already capped

by the point. However, [1]'s *DGA Status* is intended highlight TX most likely to have suspicious gassing and not as a metric for the severity of its condition. It could therefore be considered a more binary distinction between a TX seeming normal or not. In this context, it is defensible for [1] to be outputting that none of these TXs seem normal.

Fig. 4-20 shows comparatively the *Screening* output for [1] and [2] using varying sampling intervals for **TX-D**. Starting from the top row, the sampling intervals were hourly, daily, and monthly, respectively. These sampling rates were chosen as hourly corresponds to the actual sampling interval, daily appears to be highest sampling rate that [1] considered according to [1, Fig. A.6], and monthly is sufficiently infrequent as to allow the use of [1]'s 𝕋4. Fig. 4-20 shows that both approaches are heavily influenced by the sampling interval but that [1]'s *DGA Status* tends to increase with sampling interval whereas [2] tends to decrease. The first five samples for the monthly interval are not applicable for 𝕋4 and is the reason for the temporarily lower *DGA Status*; once it is applicable, the *DGA Status* rises and remains at the maximal state.

IEEE Compared to IEC: DGA Tables

Fig. 3-3 compared the limits of the tables in [1] and [2]. This subsection is focussed on the systematic differences caused by their choice of metric. Comparing table outputs in isolation requires caution, however, as their implications for *Screening* differ. Fig. 4-21 shows the outputs for **TX-A** from the limits used for [1] and [2] in the left and right columns, respectively. Values of 1 or 2 indicate remaining within or exceeding a limit, respectively. The *Score* represents the *Screening* output for that gas, where values 1–3 correspond to **L1–3** for [1], and for *Typical*, *Alert*, and *Alarm* for [2], respectively. This format shows the gases causing the overall *Screening* output. For example, it is initially CH_4, CO, CO_2, and H_2, that cause **Ł3** near the start of Y2 due to failing their 𝕋4 limits. However, it does not display the gas values or metrics relative to their limits.

For added granularity, Fig. 4-22 shows the metrics for two gases for **TX-A** for 𝕋1–4 for [1] on the left, and the equivalents for [2], L_1, L_2, G_1, and G_2, on the right. Regarding the [1]'s tables, the 𝕋3 limits are intended to represent the 95th % and thus, naïvely, approximately one in twenty samples should be flagged. For most gases, the variability in the values seem much lower than the 𝕋3 limits, except prior to Y1. CO shown in the bottom half of Fig. 4-22 is a representative example. The only period where this seems to have potentially been consistently untrue is after Q3Y3. During this period, there

was greater gassing, however, this also prompted higher sampling rates. This offers less time for gases to accumulate and effectively reduces the relative sensitivity of 𝕋3. This is demonstrated by comparing the absolute gas values of CO shown in the plot 𝕋1–2 to the metric used for 𝕋3: the latter has a much smaller incline after Q2Y3.

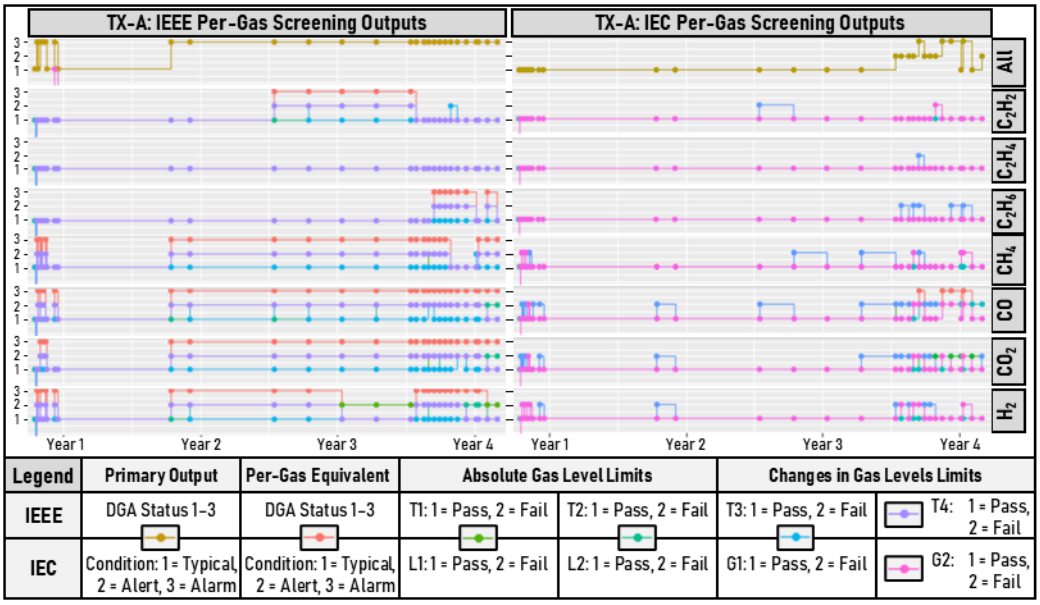


Fig. 4-21: Case Study (TX-A) IEEE and IEC Screening Table Outputs

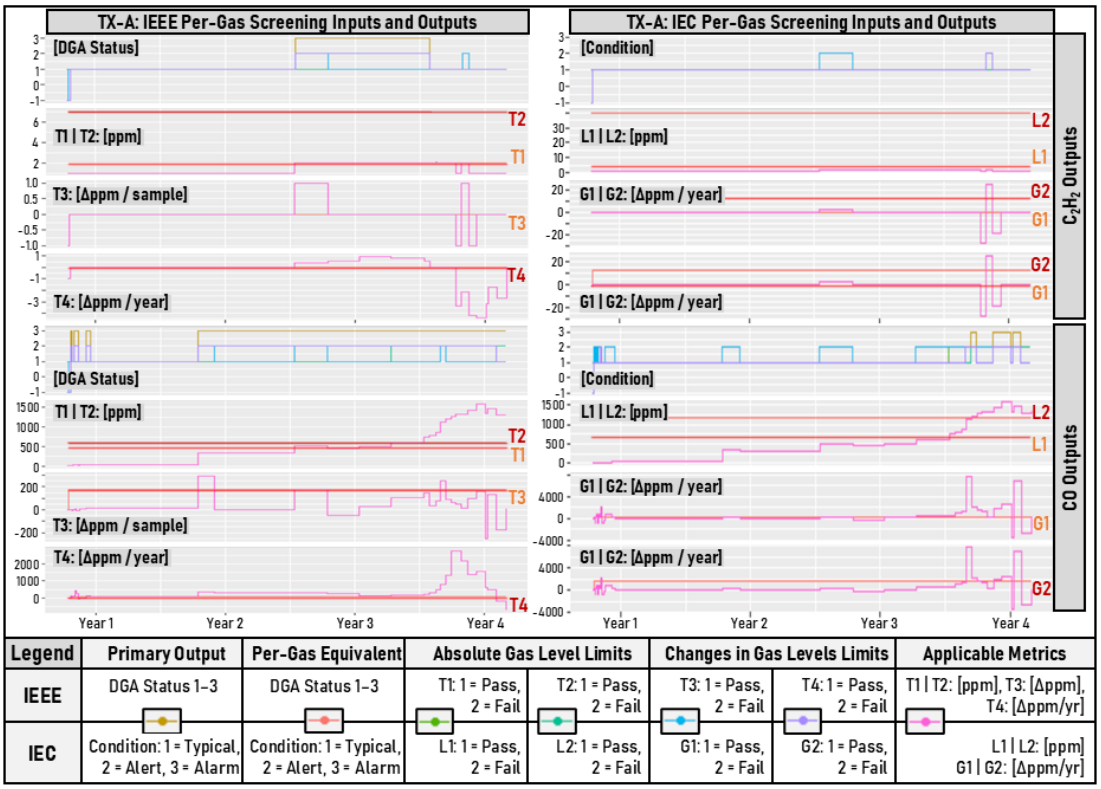


Fig. 4-22: Case Study (TX-A) IEEE and IEC Screening Table Metric Outputs for C_2H_2 and CO

This in isolation is not an issue; as previously discussed, 𝕋3 can only elevate a sample to **Ł2** whereas 𝕋4 can cause **Ł3**. Therefore, once a single sample exceeds 𝕋3, the recommended action is a *Confirmation Sample* which can then be used if necessary to calculate 𝕋4. 𝕋3–4 of [1] should be considered together, as when in isolation, they both possess major shortcomings. For example, 𝕋3 considers each delta between consecutive samples in isolation for its potential significance whereas natural intuition would suggest if multiple samples were increasing consecutively, even if in modest amounts, then there is likely gassing. This would fall under the remit of 𝕋4 and its detection of trends. However, 𝕋4 can remain invalid for initial periods until enough valid samples are present. Furthermore, during periods of increased sampling rates, it reverts to taking the first six samples, which can cause varying amplification of noise depending on the overall interval spanned. This means from approximately Q3Y3—where engineers were most concerned of potentially abnormal TX behaviour—most of the samples had an increased sensitivity to 𝕋4 and a reduced sensitivity to 𝕋3. Beyond the start of Y3 several gases eventually exceeded 𝕋2 bounds.

Considering the outputs of [2] shown in Fig. 4-21, there are fluctuating outputs prior to Y1 for CH_4, CO, CO_2, and H_2, influenced by both noise and the short sampling intervals. The column on the right of Fig. 4-22 shows the gassing rate for [2]'s 𝕋3–4 equivalents (G) oscillating, and this being amplified as the sampling intervals are shortened, indicating an increasing influence of noise. However, in Q2Y3, the gassing rate is sufficient to consistently exceed the limit for multiple consecutive samples. At this point [2]'s 𝕋1 equivalent (L_1) was exceeded, followed soon by its 𝕋2 equivalent (L_2).

There is a similar oscillation in [1]'s 𝕋3, however, it is less as it is not normalised to a year. This makes it less influenced by the sampling interval, although note its influence is in the opposite direction, i.e., a shorter interval would likely reduce the magnitude of [1]'s 𝕋3 metric whereas increase the one for [2]'s 𝕋3 equivalent. When looking at [1]'s 𝕋4, it shows a much more stable output with less dispersion about its rolling mean as was the metric's intent, though it is also less responsive as a result.

TX-D allows sampling intervals to be explored further. Fig. 4-23 shows the outputs of 𝕋3–4 for CO_2 and C_2H_6 using different sampling intervals for [1] and [2], respectively. 𝕋3–4 equivalents for [2] exhibit the same behaviour as described above and are much

larger values at smaller sampling intervals. As the sampling interval is reduced, a stronger trend is required to offset the amplified noise, else, both will tend towards an approximate 50% failure rate. [1]'s T4 at high sampling intervals, such as hourly or daily, seem an order of magnitude less effected by noise than [2]'s T3–4 equivalents, however, even this is enough to be orders of magnitude larger than the limits. For the monthly interval, when [1]'s T4 is applicable after the first six samples for **TX-D**, it frequently led to **Ł3**, as can been seen in Fig. 4-23, whereas [2]'s T4 equivalent failed less frequently. There appears to be a less pronounced impact of sampling interval on [1]'s T3, though this seems dependent on the specific gas and its respective limit. For example, T3's failure rate for C_2H_6 increased from approximately 0.7% to 1.1% to 8.0% as the sampling interval increased, with the median delta values increasing from approximately 0.0 to 0.2 to 1.6 ppm per sample, indicating an increased relative presence of the trend. This also demonstrates the relationship is tied to the absolute time interval more so than the relative time interval as there is a similar multiplicative factor between hourly and daily, and daily and monthly, but a much greater absolute difference between the latter two: gas accumulation depends on absolute time interval.

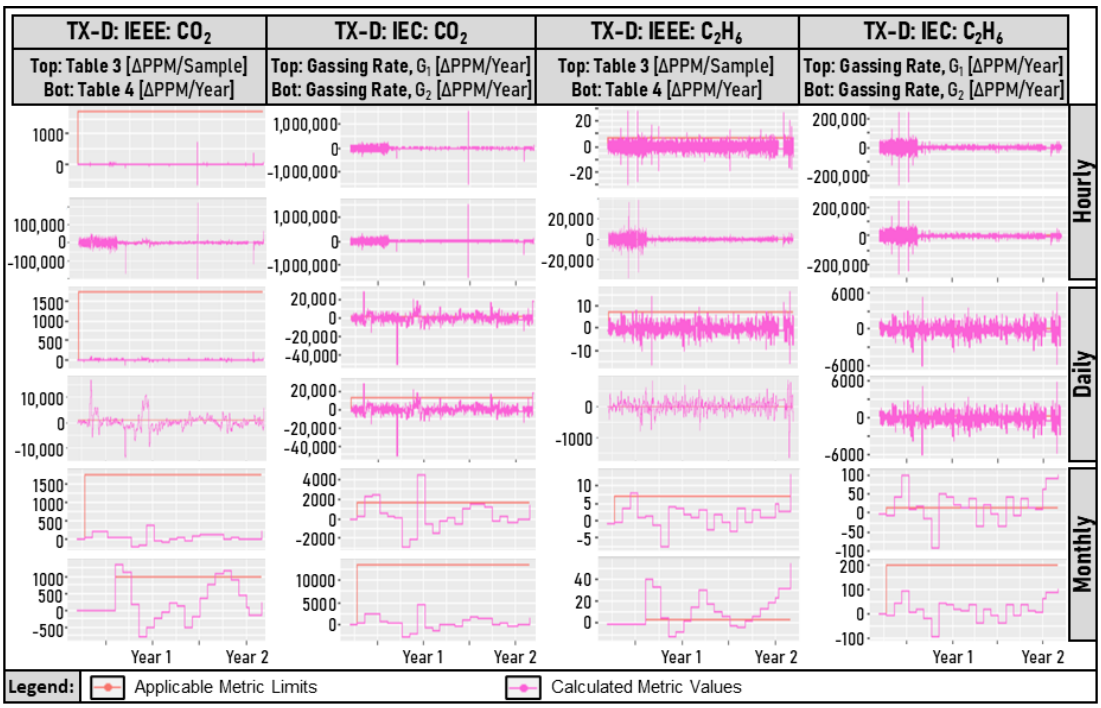


Fig. 4-23: TX-D IEEE, and IEC Equivalent, Tables 3 and 4 Metrics Relative to Sampling Rate, CO_2 and C_2H_6

The sampling intervals in Fig. 4-23 subsampled the original dataset, picking the first sample of each relevant interval. This can potentially highlight a spurious trend due to

circumstantial sample selection. Fig. 4-24 instead calculates the same metrics using all samples via pairwise selection without replacement. To emulate [1]'s 𝕋4 outputs, a simplified approach was used, where six samples were used for each calculation, and then any remaining samples forming groups of fewer than six samples were discarded. Indicatively, for CO_2, from the 6740 samples, approximately 10% of samples were discarded for the daily intervals, and 50% for the monthly intervals. Fig. 4-24 therefore, more robustly demonstrates a clear difference in behaviour between the metrics. Visually, [1]'s 𝕋3 metric remains relatively unchanged between hourly and daily intervals and modestly increases between daily and monthly intervals. In contrast, [2]'s 𝕋3–4 equivalent metric exponentially decreases as the sample interval was increased. [1]'s 𝕋4 metric seems near an order of magnitude less affected than [2]'s 𝕋3–4 equivalent metric but still shows a very large increase between daily and hourly intervals. It is somewhat comparable between the daily and the monthly intervals, but the hourly intervals seem too noisy. If basing it on [1, Fig. A.6], then arguably daily and monthly intervals should be applicable. It may suggest that if the 𝕋3 limit for [1] is applicable to daily samples, as could be interpreted based on [1, Fig. A.6], then it would be somewhat applicable to the hourly samples due to the modest difference between them. In contrast, this would also suggest that it is unlikely that a limit would be suitable for both hourly and daily, or daily and monthly sampling interval simultaneously for [2]'s 𝕋3–4 equivalents and for [1]'s 𝕋4.

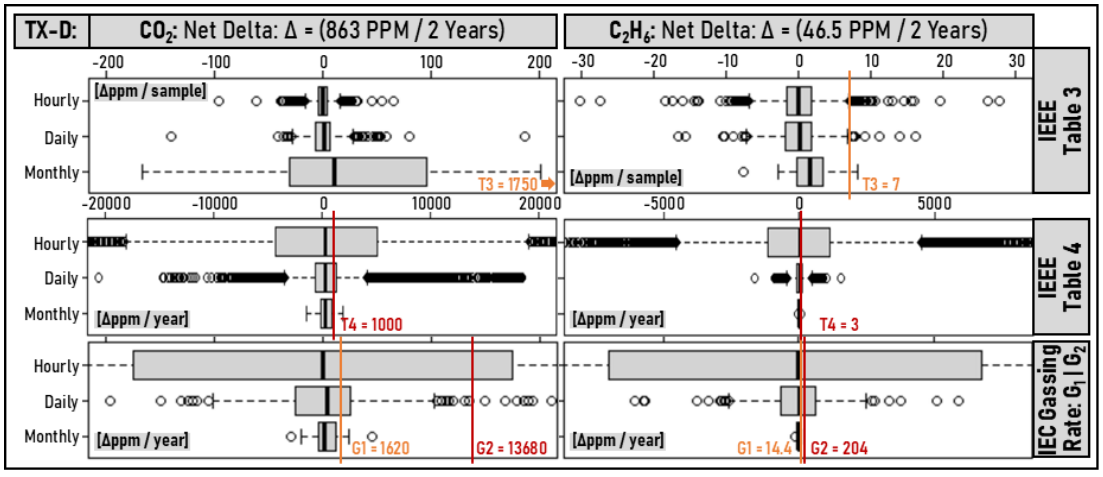


Fig. 4-24: TX-D IEEE, and IEC Equivalent, Tables 3 and 4 Outputs Relative to Sampling Rate, CO_2 and C_2H_6

However, Fig. 4-24 also raises questions about [1, p. 51]'s statements that the "typical differences between two consecutive samples (deltas) are mostly unrelated to the time between the samples". The sampling intervals discussed here are much shorter than

those intended for [1] and it may be their conclusions explicitly equivalising weekly intervals to yearly intervals are not to be extrapolated to these scenarios. However, this still draws question to [1, Fig. A.6] and whether they included the significant number of daily samples indicated to inform their limit and if it was appropriate to do so without having some stratification in its $\mathbb{T}3$ limit based on sampling intervals.

It is here argued that [1]'s $\mathbb{T}3$ is not performing as intended for OLDGA. [2]'s metrics seeks to normalise the trend, even if at the cost of amplifying noise. [1]'s metric in contrast does not normalise the trend, and consequently, increasing the sampling interval effectively decreases the measured trend. This means that a greater relative trend is required to exceed the $\mathbb{T}3$ limit. At a certain point, within the realm of expected OLDGA sampling rates, the only instances where this is realistic is during extreme gassing prior to imminent failure or simply noise. Given the latter is much more likely, the $\mathbb{T}3$ metric begins to shift from highlighting samples of interest to highlighting samples likely in need of filtering out.

It can therefore be argued to be almost as (in)effective as the [2]'s metric for ODLGA in that when it does trigger, it is almost as likely to be due to noise when at these very low sampling intervals. Its sole benefit is that it, at other times, does not flag erroneously whereas [2]'s metric is more prone to flagging approximately 50% of the samples. However, [1] can output elevated *Screening* outputs based solely on gassing rates, which is a metric arguably more affected by noise than just absolute gas levels. In contrast, [2] requires the absolute gas value to also exceed a given limit. In practice, this does little to help at high sampling rates such as for OLDGA. This issue of noise is further exacerbated by the fact that both [1] and [2] assume the worst gas as representative. Depending on the nature of the noise and whether it is correlated across the gases, this can cause a high likelihood of noise affecting at least one individual gas and thus causing an overall elevated *Screening* output.

As the sampling interval is shortened, the relevant trend is spread across more samples and is represented better by the average delta that can accumulate over time. In such circumstances, the relative importance of a metric such as $\mathbb{T}4$ in [1] increases as it can consider multiple samples. Although it also has some inherent disadvantages. For example, considering the second half of Y2 for C_2H_2 shown in Fig. 4-23, there is an

increase of 1 ppm to C_2H_2 that immediately exceeds both $\mathbb{T}3$–4. Then, even as the C_2H_2 remains at this new elevated level, the $\mathbb{T}4$ metric begins steadily climb, which may be interpreted as a worsening gassing rate, but is rather an artifact of the point shown in Fig. 4-6. Another issue with $\mathbb{T}4$ is the stipulation that it can use the six most recent samples. As the overall interval between the first and less sample is shortened, the use of multiple samples does not appear sufficient to fully mitigate noise.

However, it is important to be cognisant of cause and effect when analysing the case studies of **TXs A-C**. One could argue the primary purpose of [1] is to highlight potentially abnormal TXs to prompt closer inspection based on DGA. Given that the TX was subject to closer inspection during this highlighted period, it could be considered that the *Screening* objective was already met. In other words, had the sampling rate **not** been increased by the engineers, $\mathbb{T}3$ and/or $\mathbb{T}4$ would have flagged the samples during this region and so some of the previous critiques are irrelevant. Viewed in this perspective, [1] functioned as intended. This paradigm is consistent with Fig. 3-1 where it was stated that [1] was intended for the *Periodic Sampling Protocol* and not *Surveillance Sampling* or *Continuous Monitoring Protocols*.

This would consequently mean that another system is needed for handling TXs that are under increased sampling frequency, and arguably for those in **Ł3**. As mentioned previously, [1, Sec. 5.3.4] states that rates should be treated differently for *Continuous Monitoring*. This can cause an issue when there are many TXs at higher sampling intervals but not necessarily in a critical enough condition to warrant constant manual attention, i.e., there would remain a need for a methodology to rank and/or filter this subset of TXs. For example, Fig. 4-21 includes the overall *DGA Status* for each gas and it shows that **TX-A** would have been flagged as **Ł3** for almost 20 months before the engineers had decided to increase the sampling rate. This is unlikely to be due to an idiosyncrasy of the case study as [1, Fig. A.9] indicated that 22% of their dataset were at **Ł3**. During this period, the methodology provides no further insight as to whether the situation is worsening and is at risk of providing misleading outputs if the sampling frequency is changed. This could be argued to be the role of a *Fault Severity* assessment, but it is nevertheless an added complication. Similarly, it would mean **TX-D** cannot be directly compared to the others due to its sampling frequency.

One other practical issue with shortened sampling intervals, especially OLDGA, is the increased likelihood of desensitisation to changes. For example, an intermittent short gassing period may cause 𝕋𝟛 to fail for several consecutive samples in a day but then return to normal once the gassing stops. If the metric is reviewed on a cursory level at the end of the week, there is a risk of it assumed noise as the samples are no longer flagging. Thus, the review interval of *Screening* outputs should be considered.

There is an apparent change in behaviour coinciding with the discussed drop in H_2 levels visible only in the hourly data near Q3Y0 as seen in Fig. 4-23. This is harder to see in [1]'s 𝕋𝟛, but since [1]'s 𝕋𝟜 and [2]'s 𝕋𝟛–𝟜 equivalent metrics amplify the noise, it also highlights a marked decrease in the noise levels after this point. This highlights that even if limits are 'ideally' adjusted, the relevance must be validated periodically.

Although the intent here is not to overanalyse the specific limits, C_2H_2 is considered an outlier in that it is clearly problematic for both [1] and [2]. The implementation's 𝕋𝟛 limit for [2], and 𝕋𝟛–𝟜 limits for [1] for C_2H_2 are zero. This essentially means that unless there is a clear negative trend, it will be flagged approximately 50% of the time due to noise. There is a caveat to this as [2] states to use "<*S*" where applicable, and [1, Sec. 5] cautions against relying on low concentration, especially for *Fault Identification*, and explicitly states C_2H_2 at 1–2 ppm should be used with "particular caution". However, even when above LoD, they are still subject to some degree of noise.

The 𝕋𝟛 limits are similar between the two methodologies, but [2]'s 𝕋𝟜 equivalent limits are significantly higher than their counterpart. Although this additional buffer helps accommodate some noise, again once sampling intervals are shortened, the rate increases for [2]. This is exemplified by the C_2H_2 levels rising from 1 ppm to 2 ppm on two separate occasions. The shortened sampling interval of the second occasion caused it to exceed [2]'s 𝕋𝟜 limit. A different issue for [1] regarding C_2H_2 is that its limits of zero for 𝕋𝟛–𝟜 means any noise can cause it to flag at a maximal state of **Ł3**.

IEEE-Specific Comments

O_2/N_2 Ratio

𝕋𝟙–𝟜 in [1, Sec. 6] are stratified by O_2/N_2 ratio. It is stated in [1, p. 31] that most nitrogen-blanketed TXs examined were below the suggested limit and all air-breathing TXs were above it. However, for membrane sealed TXs, 60% were below and 40%

above, causing some uncertainty. One unclear aspect is whether the O_2/N_2 ratio is causing the differing TX behaviour or whether it is a proxy measurement of the TX type. For example, as per [1, p. 31]: "certain parameters, most notably the ratio of O_2/N_2 ... have a large influence of gases", whereas as per [1, p. 31] "the O_2/N_2 ratio was proposed for evaluation as a proxy for distinguishing sealed units from free breathing ones". The potential issue caused is highlighted in two separate examples for **TX-A**, shown in Fig. 4-25; C_2H_2 and CO_2. The highlighted period indicated when the O_2/N_2 ratio crossed the limit, causing a change in T1–4 limits. It caused a previously acceptable C_2H_2 to be flagged. For CO_2, it did the opposite, causing otherwise failing levels to not be flagged. Fig. 4-25 has four rows of plots. The top two show **L** for every gas and the combined **Ł**, first by recalculating the O_2/N_2 ratio every sample, and the second by assuming it fixed based on the first sample. The bottom two show the highlighted gases and their comparisons to the relevant tables' limits.

Fig. 4-26 shows the O_2 and N_2 levels, and their ratios for **TXs A-C**. **TX-D** did not have N_2 data so was assumed a sealed TX for the analysis and not included in Fig. 4-26. It shows during the most active gassing period between Q2Y3–Q1Y4, the O_2 levels are dropping, and that **TX-C** had the smallest drop in O_2 levels. **TX-C** also seems the least problematic in terms of gassing in general and so this is interpreted as indicative of O_2 being consumed to fuel gassing. As per [2, Sec. 5.6], the O_2/N_2 "ratio can decrease as a result of oil oxidation and/or paper ageing, if O_2 is consumed more rapidly than it is replaced by diffusion". The other results were therefore based on the assumption that O_2/N_2 should be calculated once during typical conditions and kept constant.

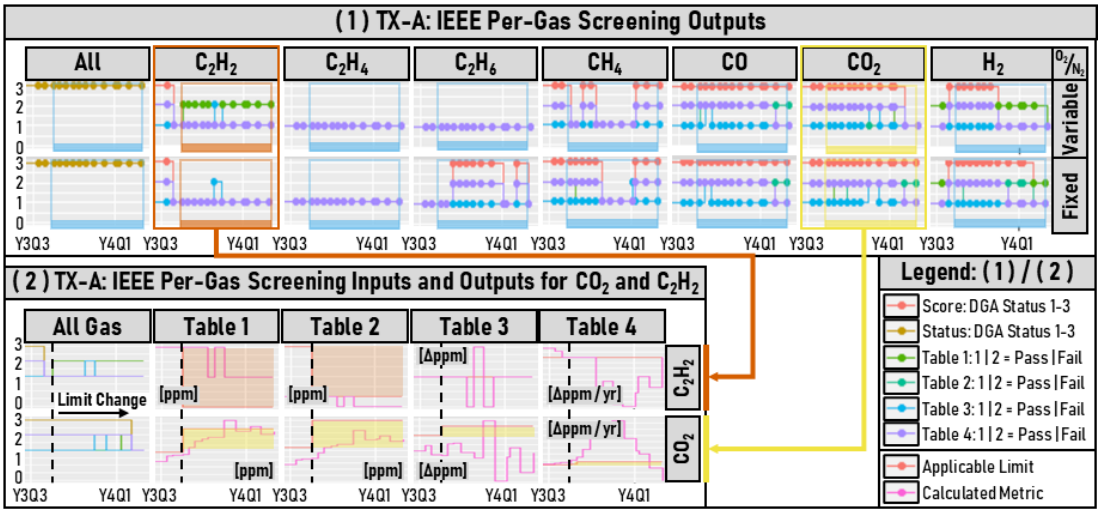


Fig. 4-25: Influence of O_2/N_2 Ratio on Case Study (TX-A) IEEE Screening Outputs

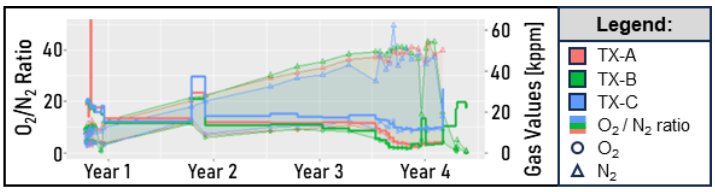


Fig. 4-26: Case Study O_2/N_2 Ratios

IEEE Resampling Procedure

There is a process for resampling for confirmation as described previously in Fig. 3-2. However, this is focussed on suspicious spikes in gas levels. Considering **TX-B** as an example, as seen in Fig. 4-16, there is a clear drop in one sample's value near Y4. This would not be flagged and in fact, the subsequent sample returning to an elevated *DGA Status* is more likely to be initially flagged as suspicious. In this example, it went from **Ł3** to **Ł1** due to this sample. In this case, the implementation could be argued to be at fault as [1, Sec. 5.1] states that if a single sample is "drastically different from earlier samples" it could indicate a sampling issue and should be confirmed via a resample. Furthermore, [1, Sec. 5.2] provides some specific guidance that could be implemented by stating "when DGA results consistently fluctuate widely (30% or two to three times the values in Table 3) from one sample to the next, it usually indicates sampling or analytical errors". Had this been implemented, then this sample would have been flagged as some of the gases dropped between two to three times 𝕋3 values and/or 30% their initial value(s). Therefore, this sample is considered invalid. Following the same reasoning, the period prior to Y1 showing excess volatility will be considered invalid sample due to unknown reasons. However, one slight issue with the phrasing used in [1] is that it poorly translates for application to C_2H_2 as its 𝕋3 limit is 0 ppm. Similarly, the alternative of 30% of its value can also unrealistic when at typically low values such as 1 or 2 ppm if the DGA has insufficient precision.

IEC-Specific Comments

Some *Screening* components essential in [2] are absent in [1], complicating direct comparison. There are three relevant components to discuss. The first is the clause that "if necessary", the deltas between subsequent samples should be used for the ratios [2, Sec. 9]. As previously discussed, this was implemented by repeating the methodology twice, once using absolute values, and once using the delta from the greater of one month prior or previous sample. Although this is mainly relevant to the *Diagnostic* outputs, [2] has a clause that states a **D2** *Diagnosis* can cause a *Screening* output of

Alert to be escalated to an *Alarm* level as shown in Fig. 3-4. This clause relating to the use of *Diagnosis* also constitutes the second component that is different to [1].

Diagnostic Override of Screening Outputs

The top-left plot in Fig. 4-27 shows **TX-A**'s *Diagnostic* outputs from [2] in a variant of its native graphical form for demonstrative purposes analogous to [2, Fig. B.1]. The bounds in Table 1 of [2], shown in Table 2-7, are shaded by colour. Both plots represent differing projections of the same three-dimensional space and should be viewed in tandem. This graphical form, however, has no inherent temporal information and can make tracking progression over time challenging. The same is true for the graphical form of *Duval Triangle 1*, shown in the top-right of Fig. 4-27. Therefore, the results will be presented in an alternative form that retains only the *Diagnosis*. The drawback is that in cases where samples are in the *Unknown* category, it is not very informative.

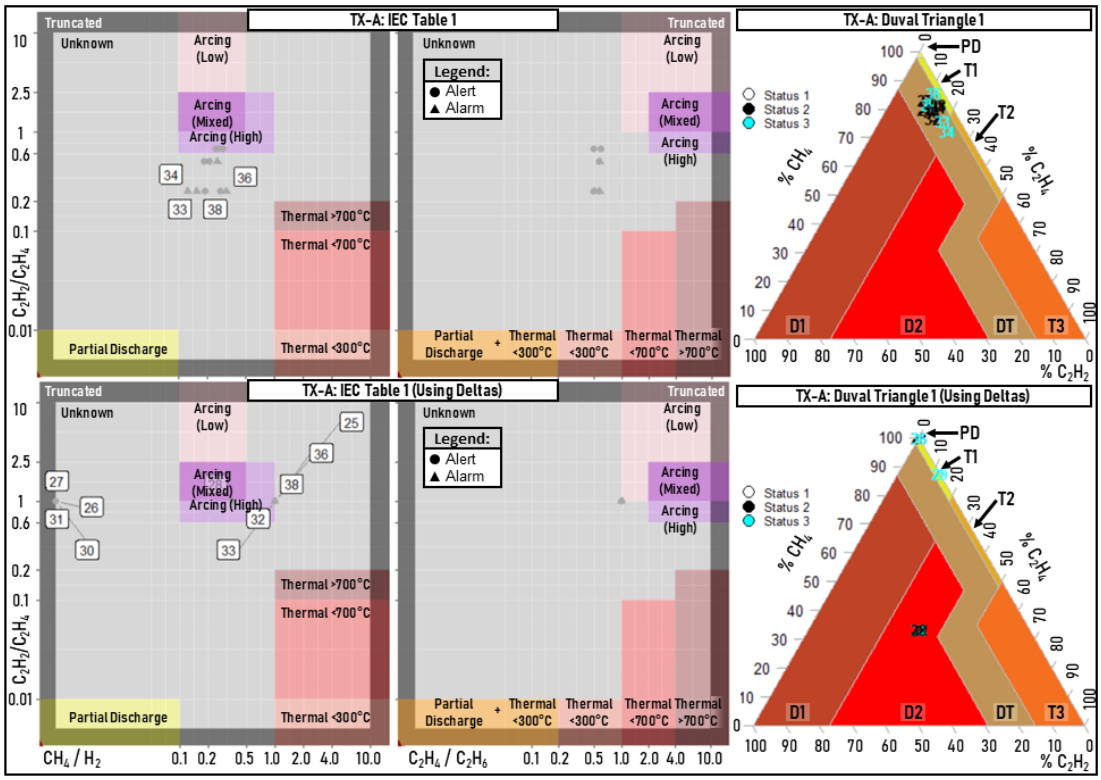


Fig. 4-27: Case Study (TX-A) IEC Table 1 and Duval Triangle 1 Diagnostic Graphical Outputs

Fig. 4-28 shows the *Diagnostic* outputs for Table 2-7 (IEC Table 1) and Table 2-8 (IEC Table 2) as applied to **TXs A-D**, plotted on the top and bottom of each paired row, respectively. Each column considers different variations that will be discussed in turn, starting with the left column, showing the outputs when using the absolute gas values. In this case, Table 2-7 (IEC Table 1) outputted the *Unknown* category for every sample

for **TXs A-C**. The ratios formed a single cluster straddling the Table 2-8 (IEC Table 2) boundary between **D**, and **D** with **PD**. Having all samples being classified as *Unknown* is clearly undesirable. For contrast, the equivalent outputs using *Duval Triangles 1-4-5* are shown in left column of plots in Fig. 4-29. During the points of elevated *Screening* outputs, *Duval Triangle 1* indicates **DT** for **TX-A**, shifting towards **T1**. **TX-B** indicates **T1** more often than **DT**. **TX-C** indicated **DT**. However, it should not be concluded that Table 2-7 and Table 2-8 simply do not work, but that as stated throughout literature, they can fail to output a categorisation for certain TXs. **TX-D** serves as a contrasting example, where Table 2-7 initially indicated *Unknown* before transitioning to **T2** and finally to **T1** near the end of the duration. Similarly, Table 2-8 indicated **T**. These align with *Duval Triangle 1* which indicated primarily **T3**.

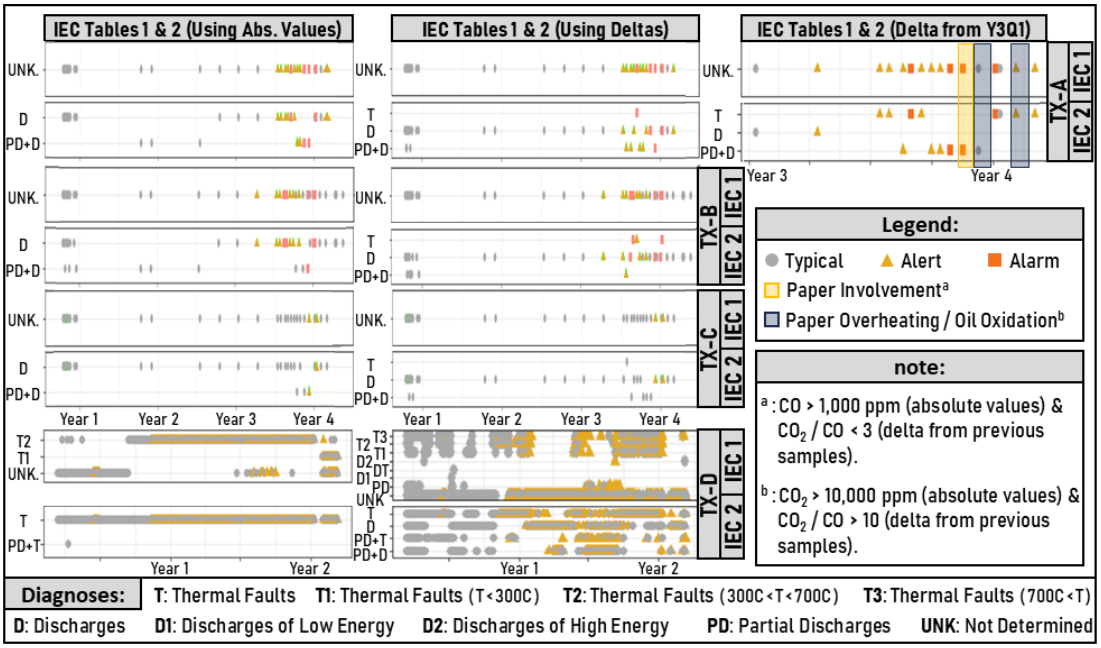


Fig. 4-28: Case Study IEC Diagnostic Outputs

Use of Deltas for Calculating Ratios

Despite *Duval Triangle 1* outputting a *Diagnosis* for any combination of ratios, it does not mean it is immune to the effects of mixed and/or transitioning *Faults*. Therefore, another approach is to use deltas to calculate ratios as previously discussed and these are shown in the bottom plots of Fig. 4-27, and the middle columns of Fig. 4-28 and Fig. 4-29, respectively. The results in Fig. 4-28 show that for **TXs A-C**, all ratios still led to an *Unknown* categorisation as per Table 2-7, though there was a slight change in outputs for the Table 2-8 method; with one or two **T** outputs whereas previously there were none. However, looking at the bottom of Fig. 4-27, the ratios are much more

scattered when taking the deltas and prone to oscillations dictated by presumably noise. This is more apparent for **TX-D** in the bottom row and middle column of Fig. 4-28, where the Table 2-7 and Table 2-8 methods outputted every possible *Diagnosis* at least once.

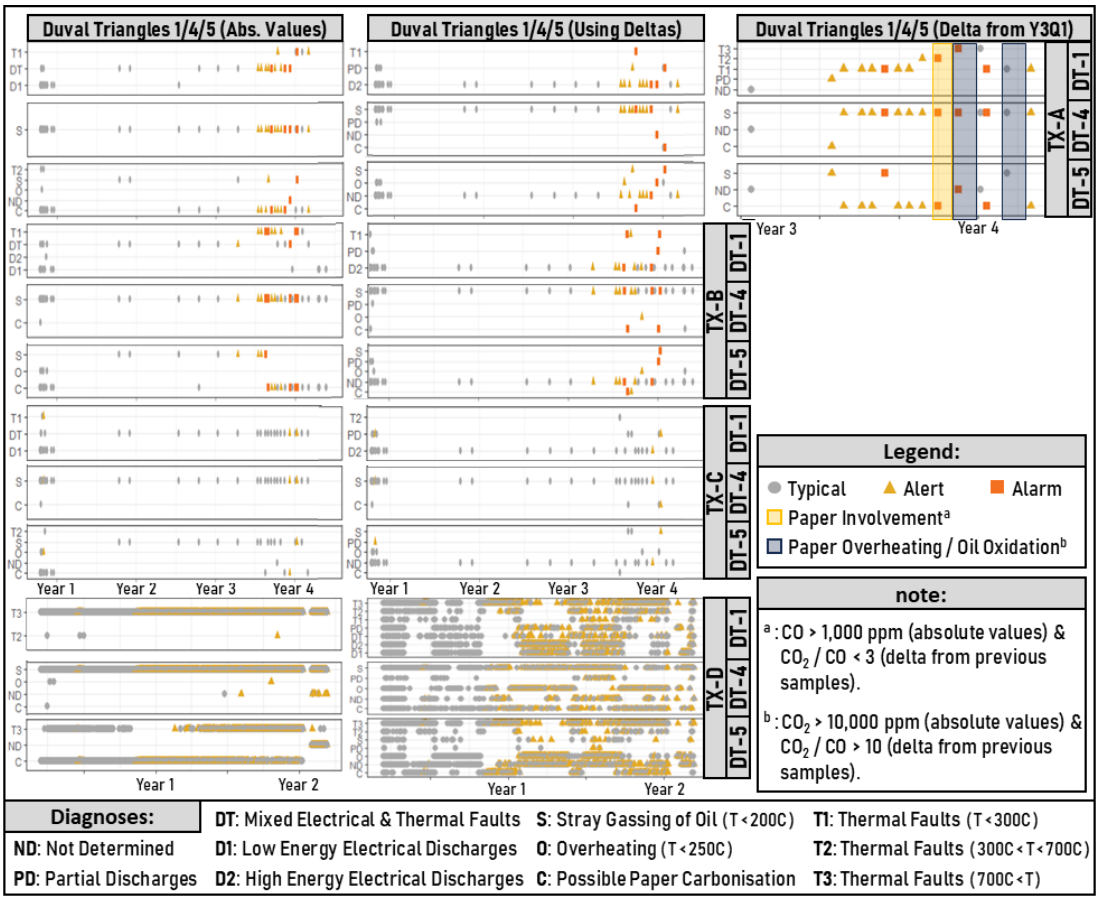


Fig. 4-29: Case Study Duval Triangles 1-4-5 Diagnostic Outputs

The middle column of Fig. 4-29, showing the outputs from the *Duval Triangles* when uses deltas, indicate a shift from **DT** to **D2** for **TXs A-C**, although the *Screening* outputs would dictate that *Diagnostic* outputs of the *Duval Triangles* should not be used for **TX-C** for all but two samples. In contrast, **TX-D** again seems heavily dictated by noise, outputting almost every possible *Diagnosis* at least once. Comparing the *Duval Triangle 1* outputs between using absolute values and deltas in Fig. 4-27 again highlights the erratic oscillations between the top of the triangle and a suspiciously repeated point in the centre of the triangle. Further inspection showed that these instances were due to deltas of either nil or negative changes, leading to a default output that problematically corresponds to **D2**—arguably the worst-case outcome. It can be concluded that a naïve

implementation of taking deltas between consecutive samples can be ineffectual and potentially misleading. This is especially the case where there is not much gassing.

This thesis is not attempting a correct *Diagnosis* of these case studies as they are unknown. However, for demonstrative purposes, if the gas levels were tracked relative to start of Y4, the right columns of Fig. 4-28 and Fig. 4-29 show the outputs for **TX-A** and shows a clear shift away from **D** and towards **T** and/or **PD**, with *Duval Triangles 4-5* indicating **S** and eventually **C**. Even in this case, Table 2-7 still outputs *Unknown* for all samples, consistently falling between **PD** and **T1** on the CH_4/H_2 axis.

Determining Paper Involvement / Carbonisation

When looking for signs of paper involvement, the absolute CO and CO_2 values were used, but the ratio was based on the delta. These are highlighted in the right columns of Fig. 4-28 and Fig. 4-29 for **TX-A**. If using CO and CO_2 values relative to the manually selected reference point of the start of Y4, no samples were flagged. Again, ground truth is not known, but comparing to Fig. 4-26, O_2 levels seem to begin to drop around Q2Y3, which is also when **T1** / **T2** / **T3** and **C** are being indicated in the *Duval Triangles 1-4-5* as shown in Fig. 4-29. As another point of reference, [1, Annex D.8]'s recommendation for indications of paper overheating or oil oxidations is similarly if there are CO_2 > 10,000 ppm but with a ratio of CO_2/CO > 20 as opposed to CO_2/CO > 10. Were this to be applied, the outputs would have been unchanged for **TX-A**. Lastly, the NEI_{paper} metric also appears to accelerate from Q2Y3 and deviate away from the N_2 accumulation rate that may be considered a baseline. Without ground truth or an extensive dataset for testing, it is difficult to draw concrete conclusions, but it is speculated that these ratios are conservative in their indications so should not be relied upon to detect all instances of paper involvement.

4.3. Conclusion

This Chapter presents analyses and findings relevant to practical deployment of the reviewed DGA methodologies: IEEE C57.104-2019 [1], IEC 60599:2022 [2], NEI [1, Annex F], and the LSA methodology [5]. Section 4.1 highlighted that the inclusion of *Uncertainty* is a nuanced and complex topic, whilst providing preliminary analyses on the expected practical significance of various factors. Section 4.2 provided case study results of automated implementation focussing on practical deployment to allow a

comparison to be made between the methodologies to inform a would-be user. Along with Chapter 3, these conclude *Research Theme 1A* by providing a detailed discussion of the impact the changes made to IEEE C57.104-2019 had on practical deployment, including a comparison with IEC 60599:2022 using real TX DGA data.

Section 4.1 Findings

The *IEC Specification* of ±15% to ±30% *Accuracy* was considered in relation to the limits found in the IEEE C57.104-2019 [1, Sec. 6]. The increased *Uncertainty* at low gas levels seemed only relevant to C_2H_2 and H_2 in the context of absolute gas concentration levels. C_2H_2 mainly due to its very low expected values and H_2 due to its much higher expected *Uncertainty* / LoD. The suggestion to include a potentially overriding minimum *Absolute Uncertainty* as per [66] was briefly explored but concluded to be only relevant to C_2H_2 at near the limit levels suggested in [1, Sec. 6]. It is stressed that this is in the context of absolute and changes in gas levels, not ratios.

As the linear regression is one of the main novelties in [1] as compared to either [2] or its own previous version, the implications of this metric was explored. The normalised delta was compared against the linear regression, where it was demonstrated how these are two fundamentally different metrics. Referring to this metric as the "**average** gassing rate" may improve clarity. The critiques of the normalised delta being heavily influenced by the time component were corroborated. Linear regression's potential to dampen *Measurement Uncertainty* was also corroborated.

The number and placement of samples when estimating gassing rate via linear regression was considered. However, drawing definitive conclusions is difficult due to the open-ended nature. Generally, increased variability requires more samples to converge on an estimate of gassing rate that would be obtained if 'all' samples were used. A primary factor for how well the estimated average gassing rate using few samples compared to 'all' samples was the extent to which the first and last sample lie on the final predicted gassing rate slope. As this would not be known, it presents a challenge. For a 'predictable' case with low variability, 3-6 samples seem to significantly influence the results and rapidly converge towards the final answer. However, for an 'unpredictable' case with high variability, it cannot be said in an unqualified sense that

6 samples can be considered sufficient to converge. The significance is that the methodology in [1] uses only up to 6 samples to calculate its average gassing rate.

Section 4.2 Findings

Four *Screening* methodologies were reviewed, implemented, and compared using real data from four TXs as case studies. The focus is the comparison of the IEEE [1] and IEC [2] methodologies. The IEEE methodology appears to be more focussed on specifically and solely providing a *Screening* output rather than it being indicative of *Fault Severity*. Furthermore, it showed a tendency to excessively flag DGA samples at its maximal level. Rather than it being interpreted as an ineffective *Screening* tool; it should be considered in conjunction with a separate second-stage focussed solely on *Fault Severity* assessment using metrics such as those in the NEI or LSA methodologies. Alternatively, it points to a need for added granularity in the outputs to aid further filtering.

Unfortunately, the coverage of the NEI methodology in [1, Annex F] lacks sufficient detail and it should be improved upon for the next edition as it seems an intuitive and promising metric. For example, it is not clear how best to use the C_2H_2 and H_2 values. Similarly, it is difficult to make use of the NEI_{paper} metric in conjunction with the carbon oxide ratios. A practical implementation based on the IEEE description alone therefore seemed incomplete compared to also using the original publications: [3], [4]. This was discussed and demonstrated further in [105], [106]. One challenge is that aspects absent in [1, Annex F] that were present in [3] and [4] are not justified. For example, the removal of the O_2-based stratification used in [4] for the applicable limits, or the exclusion of the rate of change metric used in [3].

Nevertheless, the metrics seem informative and a natural candidate for replacing older metrics such as TCG for example. The LSA metric is in some ways similar; weighting gases in accordance with perceived relevance, but it has a much greater emphasis on the ratios. This makes it very responsive and well-performing in the case studies, but also susceptible to unexpected behaviour after degassing when ratios may significantly change. For example, it was shown for **TX-D**, after the gas were reduced by over 90%, the metric increased. It should therefore not be considered a direct alternative to the NEI metric as the latter can provide helpful supplementary context. Both methods are insensitive to sampling interval, making them more applicable to OLDGA. However,

they should be considered equivalent to 𝕋1–2 from [1] and [2], and if their change over time was considered, they would likely also be susceptible to similar issues.

Regarding a comparison of [1] and [2], although both use four tables, 𝕋3–4 from [1] are substantially different. It has been demonstrated that they respond to sampling intervals differently. The case studies highlighted how [1]'s 𝕋3 reduces in sensitivity as the interval is decreased, and how [1]'s 𝕋4 is affected less by the sampling interval than the [2]'s equivalent 𝕋4 but that it is nevertheless still affected. The limits for C_2H_2 were also shown to be problematic due to the lack of tolerance for noise. Furthermore, an ambiguity in the use of the O_2/N_2 ratio in [1] was highlighted.

Although *Diagnosis* is not the focus here, the *IEC Ratio* methods did not always seem applicable to the case study examples. *IEC Ratio* method failed to provide a single output for **TXs A–C** but did give relevant outputs for **TX-D**. Similar to the *Simplified IEC Ratio*, the *Duval Triangle 1* by design provided outputs for all cases although the validity of the outputs for all methods cannot be assessed with an unknown ground truth. The intended usage of *Duval Triangles 4-5* seems vague with the potential for conflicting outputs with no guidance for resolution. The challenges associated with attempting to automate the use of changes in gas levels for ratios was highlighted as it appears very context dependent. This is true for both the *IEC Ratio* methods and the *Duval Triangles*. Furthermore, tracking the *Fault* evolution graphically was found to be quite cumbersome when there are many samples if using the original graphical formats. However, this latter point is perhaps more a critique of this thesis's implementation: adequate plot interactivity may resolve the issue in practice.

5.Proposed IEEE C57.104 Improvements

Chapter Purpose

This Chapter proposes improvements to the methodology outlined in IEEE C57.104-2019 based on the literature reviewed in Chapter 3 and the findings from the case studies in Chapter 4. This thesis primarily considers an automated implementation context, aiming to minimise subjective assessment, reserving it for critical cases. As such, the contributions are intended to improve the output granularity for easier ranking, and to incorporate a measure of *Uncertainty* for easier interpretation of the significance of the outputs.

Section 5.1 concludes *Research Theme 1B* by proposing improvements focussed addressing some of the practical barriers to deployment identified in IEEE C57.104-2019. The contributions, presented and justified through case studies, are directional suggestions subject to further refinement. The emphasis of these improvements is that they maintain the "spirit" of the methodology rather than pursuit of entirely novel approaches. These contributions are viable changes addressing *Research Theme 1B* by improving the practical performance of the IEEE C57.104-2019 methodology.

Section 5.2 addresses *Research Theme 2* by contributing a novel methodology to incorporate a measure of *Uncertainty* by extending the C57.104-2019 methodology beyond its current scope. As this is a novel topic, there is a greater emphasis on establishing the premise of the proposed methodology, including the introduction of other potentially viable candidate methodologies. This Section concludes *Research Theme 2* by exploring the relevant practical considerations for incorporating *Uncertainty* beyond the conceptual considerations covered in Chapter 3.

Chapter Structure

Section 5.1 focusses on improving practical deployment of the C57.104-2019 methodology. Sub-Section 5.1.1 recaps potential issues with the default limits in [1]'s tables and proposes improvements to the noise tolerance of the methodology. Sub-Section 5.1.2 highlights problematic tendencies of the metric designs and proposes improvements to the consistency of the methodology in the presence of a varying sampling rate. Sub-Section 5.1.3 introduces the issues regarding the derivation of the *DGA Status*. Sub-Section 5.1.4 and 5.1.5 justify improvements to the derivation of the

per-gas level *DGA Status*, **L**, and the combined *DGA Status*, **Ł**, respectively. These contributions together improve the output granularity of the methodology to facilitate the comparison of TXs in cases where they share the same *DGA Status*.

Section 5.2 considers how *Measurement Uncertainty* can be explicitly incorporated into the C57.104-2019 methodology. A mathematical background is first presented in Sub-Section 5.2.1 and then applied to both a symmetric triangular, Δ, distribution and a *Gaussian*, $\mathcal{N}$, distribution to compare the relative complexity to implement in Sub-Section 5.2.2. Sub-Section 5.2.4 presents numerical estimation as a viable methodology via either numerical integration or MC methods. Both are validated using a simple case study. Sub-Section 5.2.4 contributes two extensions to this. The first is demonstrating the integration of a *Diagnostic* stage, using the *Duval Triangle 1* as an example. The second is evaluating the impact of inter-gas correlations. Lastly, Sub-Section 5.2.5 uses the previously introduced **TX-D** data to consider some of the practical implications of various estimates. Here, the additional data afforded by OLDGA is utilised to estimate *Measurement Precision* and inter-gas correlations. The outputs using either set of assumptions are qualitatively evaluated for comparative purposes.

5.1. Improvements to Practical Deployability

5.1.1. DGA Table Limits

Chapter 4 identified that, in automated implementations of the *Screening* components, zero limits for $\mathbb{T}3$–4 can be problematic. This is relevant to the default limits suggested for C_2H_2 and can contribute to volatile outputs dominated by insignificant fluctuations driven by noise. For cases where the laboratory analysis rounds values, small noise less than the rounding tolerance may be partially filtered out, but not wholly. For example, small fluctuations about the rounding threshold will present as even larger fluctuations due to the quantisation. Furthermore, fluctuations exceeding the rounding tolerance will remain an issue. Considering the case studies in this thesis, **TXs A-C** rounded to the nearest 1 ppm whereas laboratory analysis results in Table AII and Table AIII in [71, pp. 14–15] had C_2H_2 values given to 1 decimal place.

For cases where the values are beneath the *Limit of Detection* (LoD), [2, Sec. 6] recommends replacing with "<S". In this case, if it is treated as constant, there is no issue with $\mathbb{T}3$–4, which consider changes in levels. However, in cases such as with

OLDGA which often output to more significant figures, there is then a need for either smoothing, rounding, or increasing said limits accordingly. Given that 𝕋3 is conceptually intended as an indicator for changes in gas levels exceeding normal variance whilst ignoring typical noise, it seems strange to have denoising as a prerequisite. It is therefore argued that limit values of zero should be avoided.

For example, if a single sample indicates an increase in C_2H_2 from 0 ppm to 1 ppm before dropping back to 0 ppm in subsequent samples, then currently, **Ł3** would be triggered for multiple samples due to failing 𝕋4. This is undesirable. The specific limit should be dependent on the precision of the output and can depend on both the overall duration of the included samples as well as the number of samples. In this example, depending on whether there were 3 or 6 samples within a 4-to-24-month duration, the metric output for 𝕋4 could range between 9–0.43 ppm/year, respectively. However, even a modest limit of 0.5–1 ppm/year can significantly alleviate the issue of a lingering alarm due to noise as, unless the elevated value is repeated, its impact will quickly diminish over time. Similarly, for 𝕋3, a limit for C_2H_2 of 0.5–1 ppm/sample can help. This would have no impact on the laboratory results of **TXs A-C** as they are typically at a ±1 ppm resolution, but it can help for **TX D**. Although, when considering **TX D** at its native hourly sampling rate, there remains the question of whether 𝕋3 is even conceptually valid. This is because the samples exceeding its typical variation are more likely to be due to noise than to signify a true underlying trend as discussed Chapter 4. Therefore, it is meant more in a hypothetical sense; that data similar to **TX D**, where the values have not been rounded to the nearest 1 ppm, would benefit from the slightly elevated limit. Even in the hourly sampling context, where 𝕋3 is of dubious relevance, having a higher limit to flag less is beneficial as it less often interferes with the output.

5.1.2. DGA Table Derivations

As highlighted in the results of the case studies (Fig. 4-23 and Fig. 4-24), [1]'s 𝕋4 metric has a similar susceptibility to noise as [2]'s metric for its 𝕋3–4 equivalents once the sampling rate exceeds once per month, albeit to a lesser severity. This can occur primarily under three scenarios. The first is due to the use of OLDGA and its typically high sampling rates. The second is if a TX has a relatively high CoF and is thus subject to routinely high sampling rates. The third is if a TX is considered to have a relatively

high PoF perhaps due to suspicious values from previous samples and is now subject to increased sampling rates.

It is stipulated in [1, Sec. 6] that 3–6 samples between 4–24 months are required for 𝕋4. If more than 6 samples are available within 24 months, then the latest six are used. **TXs A-C** had increased sampling rates once there was suspicion of an issue, leading to 6-sample intervals dropping to approximately 10 weeks. Though this demonstrates how the third scenario can actualise, the increase in sampling rate was not as severe as could otherwise be. For example, [1, Fig. A.7] indicates approximately 20% of samples were taken daily. Therefore, **TX-D** is instead used to illustrate the potential impact of the issue. Fig. 5-1 shows the impact of shifting from monthly to daily sampling for CO_2 for **TX-D** in the left column as a demonstrative example. In this example, only 𝕋4 is driving an elevated *DGA Status* for CO_2. The initial duration is inapplicable until there are sufficient samples which take longer for the monthly sampling rate to achieve as compared to the daily sampling rate. However, this aspect is here being considered irrelevant to the discussion and the focus in instead on the remaining duration. The results show that the 𝕋4 metric is almost unrecognisable for the daily sampling if being compared to the monthly sampling. Similarly, the daily sampling's **L** is much more often elevated due to 𝕋4 failing more frequently.

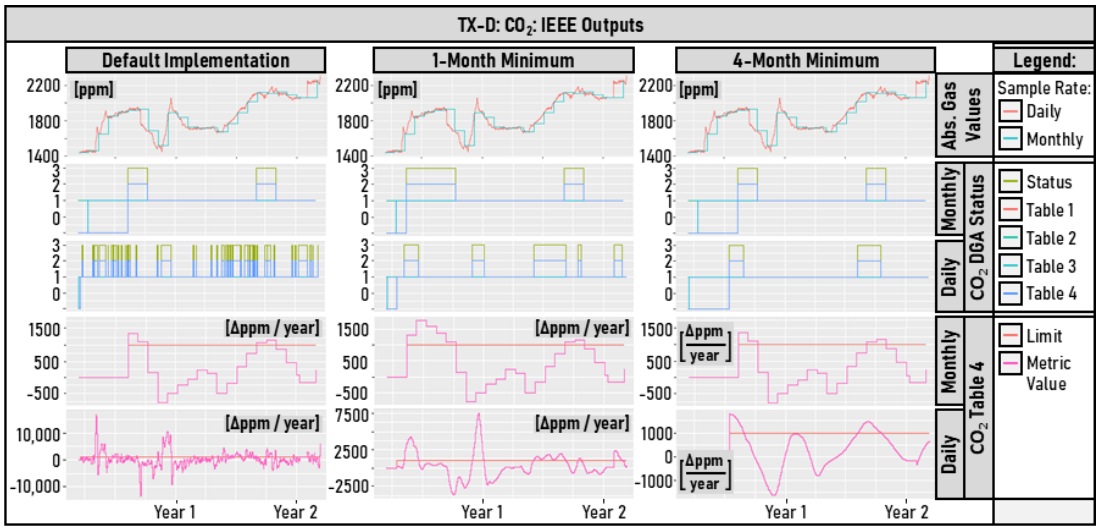


Fig. 5-1: Case Study (TX-D) IEEE Table 4 Metric Outputs for CO_2 Relative to Sampling Rate

One could argue this can be alleviated by adding an additional category to 𝕋4 for shorter durations that may be more applicable. However, [1, Fig. A.7] indicates the data for this was already available and this was not pursued. Furthermore, [1, Annex B]

emphasises the impact that the overall duration has on the metric such that noise can begin to dominate over the actual gassing rate. The premise for using a linear regression of multiple samples was to alleviate this, however as Section 4.1 showed, there is no reason to assume that 6 samples is sufficient to converge onto the 'true' value. Furthermore, the effect of shortening the overall interval seems somewhat exponential and therefore perhaps necessitating multiple additional categories.

Within the context of the effect of transitioning between sampling rates, the current ruleset of using the latest six samples is insufficient to safeguard against the potentially distorting effects it can have. Instead, it is proposed that the simple modification of prioritising and enforcing a minimum duration rather than the maximum number of samples can improve performance for this scenario. The second and third column of Fig. 5-1 shows two candidate variants, using a 1-month and a 4-month minimum, respectively. The 4-month minimum is selected as it mirrors what is already mentioned in 𝕋4 and is therefore also more readily comparable to the limits contained in 𝕋4. The 1-month minimum is somewhat arbitrary, based on engineering judgement. As previously mentioned, [3, Sec. F] states, monthly values are "an intuitively good time scale for expressing gas accumulation rates in transformers". This duration will only trigger if the sampling frequency is more than weekly. This shorter minimum should also mean it is more responsive to changes than the alternative 4-month minimum at the expense of being more susceptible to noise and potentially requiring another limit category for 𝕋4.

Regarding the relevance of such a change, [1, Sec. 6.1] states that **Ł2–3** could be considered grounds for increased sampling frequencies. [1, Fig. A.9] shows that their dataset had 21% of samples at **Ł2** and 22% at **Ł3**. This indicates that approximately 20–40% of samples were either triggering, or already under, increased sampling frequencies. Therefore, this scenario is relatively common, and so, having a *Screening* methodology that can accommodate it should be considered important.

It is not the intent here to propose and defend a specific minimum duration, but rather to argue that one should be in place. Ideally, the value would be based on a specific dataset. When comparing the outputs in Fig. 5-1, enforcing a 4-month minimum duration leads to a recognisable trend between the two sampling rates and similar

outputs for **L**. Predictably, a 1-month minimum leads to results that are somewhere in between having no minimum and having a 4-month minimum. Although **L** changes significantly when the sampling rate is changed, it is not implied the results are less valid per se; as the top row of Fig. 5-1 shows, there **is** a seemingly rapid increase in gas levels towards the start of Year 1 that may well warrant flagging. The key aspect is attempting to minimise the impact caused by changing the sampling rate such that it does not become the reason samples are being flagged. Within this case study, a 4-month minimum achieved the desired goal.

5.1.3. DGA Status Derivations Preface

There are two motivations for the topics being discussed here. The first goal is to maximise the informativeness of **L** and **Ł** for *Screening*-related decision-making. The second goal considers a broader context than DGA. As discussed in Chapter 1, a TX is a multi-component asset that can be subject to multiple *Condition Monitoring Techniques* concurrently. The current 1–3 scale can complicate its integration into a broader *TX Assessment Index* (TAI) incorporating multiple outputs. There is therefore an incentive to first standardise it to a more conventional 0–1 scale.

Given that it is a *Screening* index that is intended to highlight TXs that seem unusual rather than pass judgement regarding *Fault Severity*, one can argue its current very discrete nature is a non-issue. The intent is to increase granularity for when too many TXs are flagged, whilst retaining its decisive nature. The derivation of **Ł** is based **L**, as was described in Chapter 3. Therefore, each will be explored independently below with the goals of increasing output granularity and rescaling to a 0–1 index.

In lieu of specific case studies, all potential output combinations for **L** are considered to demonstrate the entire range of potential impacts of the proposed modifications.

5.1.4. Per-Gas DGA Status

The intent is to increase granularity of the *Screening* output. As published in [75], the proposed modification is conceptually simple: given that $\mathbb{T}1$–2 use the same absolute gas level metric, they are here considered to be two ends of a scale. Currently, $\mathbb{T}3$–4 aside, values less than the $\mathbb{T}1$ limit are equivalent to **L1** and those more than the $\mathbb{T}2$

limit are equivalent to **L3**. Values in between are **L2**. The modification linearly interpolates between the $\mathbb{T}1$–2 limits to provide outputs between **L1–3**, exclusively.

The modification is intended to be implemented alongside a rescaled *DGA Status*, where **L1** ≙ 1, and **L3** ≙ 0. Thus, **L2** ≙ 0.5 would be an expectation. However, in this implementation, it would instead scale between 0–1 depending on how close the absolute gas levels were to a given limit, as shown in Fig. 5-2 using C_2H_4 as an example.

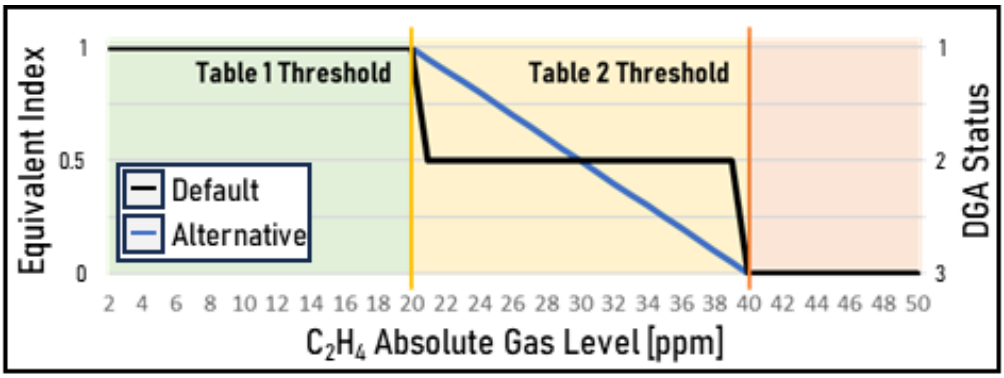


Fig. 5-2: Example Impact of Linearly Interpolating Screening Output

$\mathbb{T}3$–4 are treated differently. They are of different units and are intended to capture different aspects and so cannot be paired in the same way as $\mathbb{T}1$–2. Perhaps including an additional table to mimic a similar system as [2]'s methodology, where it has two tables for gassing rate, could allow for a similar approach as for $\mathbb{T}1$–2. This was not pursued here. Instead, $\mathbb{T}3$ is considered to represent the lower limit above which samples have undergone potentially significant changes and thus will have no impact if not exceeded. Similarly, $\mathbb{T}4$ is argued to also be more affected by noise than absolute gas levels and that its limit represents the lower bound of significance. Therefore, unless exceeded, it too will have no impact. If either are exceeded, they are treated largely as in the original methodology: failing $\mathbb{T}3$ would result in **L2** ≙ 0.5 and failing $\mathbb{T}4$ would result in **L3** ≙ 0. The minor modification here is that if the outputs from $\mathbb{T}1$–2 would otherwise result in **L** < 0.5, it would override the outputs from $\mathbb{T}3$, i.e., the worst case will be selected from the outputs of the tables to represent **L**.

These modifications should not compromise the methodology's original intent as only **L2** is granulised without impacting other outputs. However, the added complication of calculating the linear interpolation between $\mathbb{T}1$–2 may be considered a downside.

5.1.5. Combined DGA Status

Overview

Using the worst-case **L** to represent **Ł** may result in excessive information loss. While it could be argued the added information is unnecessary for deciding to increase *Surveillance* based on the *Screening* outputs, it can be included with minimal disruption whilst having foreseeable benefits. Sub-Section 2.2.4 discussed calculating a TAI and introduced Table 2-2 of [7, Sec. 2.4] which suggested a range of potential approaches for aggregating outputs. While eight distinct techniques were suggested, many are interchangeable through trivial alterations. Therefore, only the following are discussed:

Count per Category

In the context of the *DGA Status*, this could either be a count of **L** or the count of the gases failing $\mathbb{T}1$–4. Perhaps the main issue is that it does not address how the TXs should be ranked and thus still requires a second stage of compression. This approach represents presenting the maximum information but minimal compression.

Worst case approach

This is [1]'s current implementation and represents maximal indicator-compression. This is stated as not only the simplest approach in [7, Sec. 2.4], but also as a "transparent" one. Although its derivation may be transparent in that it is easy to understand, it is not transparent as a metric as it can mask both the number of worst-case indicators, as well as all other less severe indicators. The result is that it is prone to large jumps in output without warning of escalations in the interim periods. An alternative variant is suggested in [7, Sec. 2.4] that outputs the worst-case score alongside a count of its occurrences to help improve transparency.

Summed or average scores

Perhaps the most intuitive implementation would be summing **L** to give **Ł**. In its simplistic form, this is inadequate as the recommended action is primarily motivated by the worst-case gas. This is consistent for both [1] and [2]. A single gas at a **L3** may be seen as more severe than all gases at a **L2**, for example. Another potential minor issue is that not all gases are present for all TXs, and a summing approach can be unduly affected by the number of gases. Lastly, this approach also does not lead to a 0–1 scale

and would need to be subsequently normalised. This normalising would make both summing and averaging equivalent.

Weighting can help differentiate between the different scenarios. One avenue to explore is considering the relative importance of each gas to weight. For example, C_2H_2 may be considered a higher priority than CH_4. However, it is argued inappropriate as the *DGA Status* is intended as a *Screening* output and not as a *Fault Severity* output. Therefore, relative to their respective expectations, each gas can be of equal 'unusualness'. Although, there is some nuance regarding gases such as H_2 which may be argued as inherently less reliable. This may be used as a basis for underweighting them, this is difficult to quantify and not applicable to most gases.

Another approach is to weight each level of **L** differently to create a non-linear scale. There are of course many ways to approach this depending on the desired properties of the output. In general, there are trade-offs to balance between the complexity, ease-of-interpretation, and granularity. This will be explored further shortly.

Hybrid score

This is simply a mixture of the above. An example provided in [7, Sec. 2.4] shows the summed score alongside a colour representing the worst-case score. This would therefore give two prospectives and further insight. For example, one could then sort by worst-case score initially, then do a secondary sort based on the summed score for a more granular output. This arguably provides 'the best of both'. However, to fulfil the function of condensing the original outputs, not everything can be included.

Comparisons

The intent is to augment the *DGA Status* in a manner that does not significantly disrupt its current interpretation. Therefore, it is considered a requirement for the worse-case **L** to be distinguishable for all circumstances, whilst also improving granularity, and scaling between 0–1. For example, Fig. 5-3 shows the output of various metrics for all potential combinations of **L** of a TX assuming it has available all gases where the default implementation shown in black. The values along the abscissa represent a count of **L1–3**, where the 100s signify the count of **L3**, the 10s signify the count of **L2**, and so on. The shaded regions represent if the previously discussed **L** modifications were implemented—these are addressed later. An ideal metric would minimise repeated

values and clearly convey the worst-case **L** to preserve the intent of the original approach.

Equation (59), based on Equation (1) from Section 2.2 is shown in blue.

$$TAI' = \left(T\hat{A}I - TAI\right)/T\hat{A}I, \tag{59}$$

where the TAI is defined in Equation (1), and $T\hat{A}I$ represents the maximum possible TAI. This variant simply scales the output between 0–1. The parameter values used for k and i for Equation (1) were 3 and 8, respectively. i was described as "equal to or greater than the number of failure modes included" and so were applied here as equivalent to the number of gases, requiring at least a value of 7 in this case [7, Sec. 2.4]. This is equivalent to summing each **L** where **L1** ≙ 1, **L2** ≙ 8, and **L3** ≙ 64.

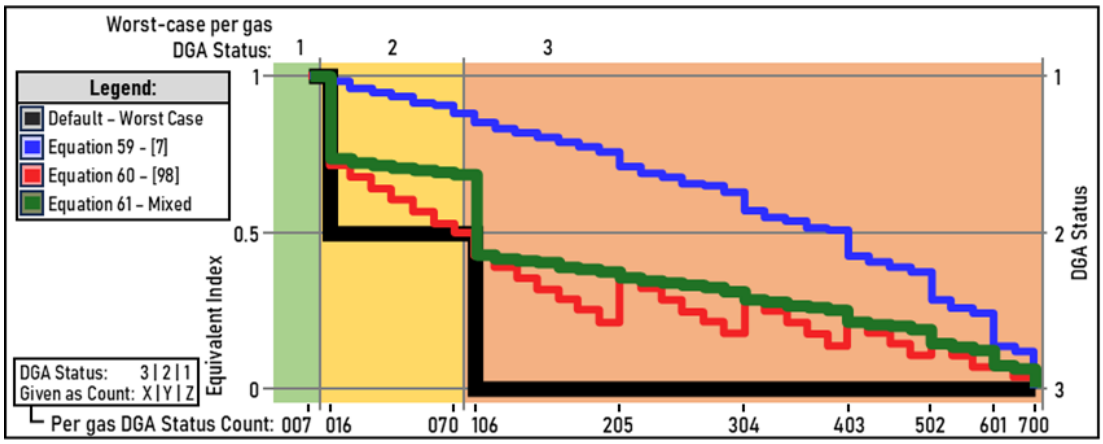


Fig. 5-3: Comparison of Differing Combined DGA Status Derivations

There is a subjective nature to evaluating these metrics. Given the stated goal to emulate the heavy emphasis on the worst-case **L** as in the original, the output given by Equation (59) fails in this regard. If this is considered an arbitrary constraint intending to reflect the original implementation only superficially with no practical benefits, then Equation (59) exhibiting high granularity across the spectrum may be considered a suitable alternative. Otherwise, [75] published a different approach as shown in Equation (60) and in Fig. 5-3 as the red line.

$$TAI = \left(\check{L}_g + \bar{L}_g\right)/2, \tag{60}$$

where $\check{L}_g$ and $\bar{L}_g$ are the worst-case and average **L**, respectively. The terminology can be confusing as the original *DGA Status* ascends with severity whereas the 0-1 scaling descends. In this case, the 0–1 scaling is being used, and the equation is simply the average of the average of all **L** and the worst-case **L**. This creates a new dynamic that

more closely matches the original implementation by having large differences between scenarios of differing worst-case **L**. A potentially trivial property it also possesses is that it matches the original methodology's equivalent values in cases where all **L** are the same, i.e., at values 1, 0.5, and 0. One shortcoming of the approach is that within these scenarios, there are repeated values which can create some ambiguity. Another potential issue is that it will weigh the same given **L** differently depending on the other values for **L**, in that the first instance of **L3** will have a greater impact on the overall score than subsequent **L3**. Although, this also applies to the original implementation.

The final considered alternative is a combination of what was originally published in [75] and Equation (59) such that there are no repeated values and a clear distinction in values depending on the worst-case **L**. This simply replaces how $\bar{L}_g$ was calculated in favour of the calculation method of Equation (59). This is shown in Equation (61) and as the green line in Fig. 5-3.

$$TAI = \left(\check{L}_g + TAI'\right)/2. \tag{61}$$

This final alternative loses the parity with the original under circumstance of all being **L2**, but instead has no repeating values, which is argued more important. A potential limitation is the reduced granularity as compared to Equation (60) in scenarios where the worst-case is **L2**.

Again, as with the other proposals, the intent is more highlighting a potential weakness in the current implementation and a general direction towards improving it rather than imposing a definitive solution. Any of the three latter proposals are improvements to the original implementation and can be adopted with minimal disruption. For example, if using Equations (60) or (61) and seeking to map back to the original implementation, values of 1, 0.5–1, and ≥0.5 would equate to values of 1, 0.5, and 0, respectively. The values for Equation (59) will be slightly less intuitive as it will depend on the number of gases included, but for the expected 7 gases, the values for the thresholds would be 1, 0.86–1, and ≥0.86, respectively.

Integration with Per-Gas DGA Status modifications

If combining this change to the derivation of **Ł** with the previous change to the derivation of **L**, the thresholds will differ as **L2** will be defined by a non-inclusive 0–1 range as opposed to the fixed 0.5 value. This means the integer counter n going from

0–2 for **L1–3**, respectively, for Equation (1) needs adjusting. Instead, where there would have been **L2**, n is equated to two minus twice the new **L2** value such that n could range 0–2 non-inclusively in these cases. The shaded regions in Fig. 5-3 indicate the potential range of values for each respective metric. This range represents the worst-case scenario where **all L2** values are practically equivalent to **L3**, with an assumed value of 0.01, and the best-case scenario where they are **all** assumed a value of 0.99. For instance, a scenario shift from '160' to '205' can be practically equivalent to a shift from either '700' to '205' or from '106' to '205', depending on the assumed value of **L2**. This causes the seemingly erratic ranges along the y-axis shown in Fig. 5-3. In practice, it would be very unlikely to have **all L2** at either end of this range. Thus, it is neither a priority nor practical to pursue a metric with no overlap when integrating the **L** modifications suggested in Sub-Section 5.1.4.

5.2. Integrating Measurement Uncertainty

The final topic is the explicit inclusion of *Measurement Uncertainty*. As discussed in Chapter 3, DGA samples have finite *Accuracy*, which may present as exceeding a limit whereas their true value would have them beneath it. Three variants of [1], **S1–3**, aligned with its *Protocols*, will be discussed as this both comprehensively addresses [1] and offers a natural progression of increasing complexity. The simplest, **S1,** is the *Initial Sampling Protocol* which requires only $\mathbb{T}1$–2 to calculate **L**. The second, **S2**, is the *Periodic Sampling Protocol* using $\mathbb{T}1$–3. Lastly, **S3** is the *Periodic Sampling Protocol* using $\mathbb{T}1$–4. Unless stated otherwise, the following will assume the original implementation of [1] as explained in Sub-Section 3.1.2 without the previous modifications. The general mathematical representations for $\mathbb{T}$, **L**, and **Ł** will be discussed first, then practical methods for calculating the relevant outputs will be explored. Numerical integration and a MC simulation are the two explored implementations. Lastly, an example using **TX-D** demonstrates an implemented IEEE C57.104 methodology [1] with integrated *Measurement Uncertainty* propagation.

5.2.1. Theoretical Background

It is assumed that an unbiased unimodal distribution can represent the *Uncertainty* of a gas measurement. For the discussion regarding algebraic solutions, there are two assumptions of independence. The first is that for a given gas, samples are considered

serially independent. The second is that for a given sample, the gases are considered independent of one another. The work presented here utilises only the *Accuracy* metric. Please refer to Chapter 3 for a more detailed discussion on these assumptions.

Combined DGA Status

Equations (62)–(64) show the probability of **Ł1–3** derived using **L1–3**, respectively.

$$P(Ł = 1) = \prod_{g=1}^{k} P(L_g = 1), \tag{62}$$

$$P(Ł = 3) = 1 - \prod_{g=1}^{k} 1 - P(L_g = 3), \tag{63}$$

$$P(Ł = 2) = 1 - P\big((Ł = 1) \cup (Ł = 3)\big), \tag{64}$$

where k represents the total number of gases, g, included in the DGA sample that have relevant tables associated with them. Ł and L represent **Ł** and **L**, respectively. Although there are three *Protocols*, **S1–3**, that will be discussed, the derivation of **Ł** is the same.

Per-Gas DGA Status

As described in Fig. 3-2, there are three relevant *Protocols*, **S1–3**, within [1]: the *Initial Sample*, *Periodic Sampling* using 𝕋1–3, and *Periodic Sampling* using 𝕋1–4.

S1: Initial Sampling Protocol, (𝕋1–2)

For **S1**, only 𝕋1–2 are needed to calculate **L1–3**, as shown in Equations (65)–(67), where $\neg T$ and T represent exceeding, and not exceeding a limit for a given table for a gas denoted by the subscript g. respectively.

$$P(L_g = 1) = P(T_{1,g}), \tag{65}$$

$$P(L_g = 2) = P(\neg T_{1,g} \cap T_{2,g}), \tag{66}$$

$$P(L_g = 3) = P(\neg T_{2,g}). \tag{67}$$

S2: Periodic Sampling Protocol, (𝕋1– 3)

S1 is a niche case, and most samples are instead expected to be in **S2–3**. For **S2**, only 𝕋1–3 are required to calculate **L1–3** as shown in Equations (68), (69), and (67).

$$P(L_g = 1) = P(T_{1,g} \cap T_{3,g}), \tag{68}$$

$$P(L_g = 2) = P\left((T_{1,g} \cap \neg T_{3,g}) \cup (\neg T_{1,g} \cap T_{2,g})\right). \quad (69)$$

S3: Periodic Sampling Protocol, (𝕋1–4)

S3 requires 𝕋1–4. 𝕋4 generally requires 3–6 samples over 4–24 months. Equations (70), (71), and (72) show how the tables are used to calculate each **L1–3**, respectively.

$$P(L_g = 1) = P(T_{1,g} \cap T_{2,g} \cap T_{3,g} \cap T_{4,g}), \quad (70)$$

$$P(L_g = 2) = P\left((\neg T_{1,g} \cap T_{2,g} \cap T_{3,g} \cap T_{4,g}) \cup (\neg T_{3,g} \cap T_{2,g} \cap T_{4,g})\right), \quad (71)$$

$$P(L_g = 3) = P\left((\neg T_{2,g} \cap T_{4,g}) \cup (\neg T_{4,g})\right). \quad (72)$$

Marginal Distributions of Tables

The outputs from each relevant table are combined to calculate **L**. Each are discussed.

Absolute Gas Values, (𝕋1–2)

𝕋1–2 use the gas value for the current sample, Y_1, and the probability of it being within their respective limits, τ, are shown in Equations (73) and (74). The subscript g is henceforth omitted to avoid clutter.

$$P(T_1) = P(Y_1 < \tau_1), \quad (73)$$

$$P(T_2) = P(Y_1 < \tau_2), \quad (74)$$

where the subscript for Y denotes the sample order relative to the newest sample, ascending in value where 1 represents the most recent sample. Using the *Cumulative Distribution Function* (CDF) appropriate for Y_1, makes these trivially solvable.

Delta in subsequent Absolute Gas Values, (𝕋3)

𝕋3 is the difference between two consecutive samples, and the probability of it being within its limit is shown Equation (75). This represents an equivalent double integral with an outer integral across the domain between $\check{Y}_1$ to $\hat{Y}_1$, and an inner integral along the range of Y_2 at each point, as shown in Equation (76). f_i is the relevant *Probability Density Function* (PDF) for a given sample, Y_i. It may be simpler to use the predefined function of the inner integral of Y_2, F_2, in lieu of the second integral as shown in Equation (77).

$$P(T_3) = P(Y_{1-2} < \tau_3). \quad (75)$$

$$P(Y_{1-2} < \tau_3) = \int_{\check{Y}_1}^{\hat{Y}_1} f_1(y_1) \times \int_{\check{Y}_2}^{\hat{Y}_2} f_2(y_2)\, dy_2 dy_1, \tag{76}$$

$$P(Y_{1-2} < \tau_3) = \int_{\check{Y}_1}^{\hat{Y}_1} f_1(y_1) \times [1 - F_2(\check{\gamma}_2)]\, dy_1, \tag{77}$$

where γ represents the appropriate limit. For example, $\check{\gamma}_2$ will be the greater of $\check{Y}_2$ or $y_1 - \tau_3$. Where the overriding term defining the limit changes, the integral must be split and done by parts.

Average Gassing Rate, (𝕋4)

The metric for 𝕋4 is the slope coefficient, β_1, obtained via a linear regression. Section 4.1 discussed some of the derivations, in particular, Equation (55) demonstrated the additive nature, which can also be expressed as in Equation (78), from [47]:

$$\bar{\beta}_1 = \sum_{i=1}^{n} \frac{\Delta\beta_1}{\Delta y_i} \bar{y}_i, \tag{78}$$

$$\frac{\Delta\beta_1}{\Delta y_i} = (\delta_{ij}, \delta_{ij}, \dots, \delta_{ij}), \tag{79}$$

where $\Delta\beta_1/\Delta y_i$ is defined in Equation (79) and represents the change in the slope coefficient as y_i is changed. For brevity, $\Delta\beta_1/\Delta y_i$ is henceforth termed c_i. δ is the Kronecker delta function [102] defined in Equation (80).

$$\delta_{ij} = \begin{cases} 1 & if\ i = j, \\ 0 & otherwise. \end{cases} \tag{80}$$

Please note that Equation (79) is only varying the sample values and not the sample times. The sample times are relative such that the oldest sample is $x_N = 0$; this creates a positive slope for increasing gas levels over time. Another aspect assumed trivial is the appropriate scaling such that β_1 represents the units of ppm/year. The minimum and maximum values of β_1 are calculable via Equations (81) and (82). Of course, for normal distributions, this represents an unbound range from $-\infty$ to ∞.

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i \times \hat{h}(y_i), \tag{81}$$

$$\hat{h}(y_i) = \begin{cases} \check{y}_i & if\ x_i < \bar{x}_i, \\ \hat{y}_i & otherwise, \end{cases} \tag{82}$$

where the function, $\hat{h}(y_i)$, picks the worst-case value to maximise β_1. The minimum value would be evaluated using the equivalent function, $\check{h}(y_i)$, which would simply reverse the logic used in Equation (82).

Then to determine the PDF of β_1, denoted by g, the equivalent of summing, convolution, is required. The convolution operation, $*$, is defined generically by Equation (83). The notation, $\circledast$, here represents repeated convolutions analogous to the summation operator, Σ, and thus Equation (84) shows the PDF of β_1, denoted by g_N. Its integral is then defined by the function G_N as shown in Equation (85) representing the CDF of β_1.

$$(g_i * g_j)(y) = \int_{-\infty}^{\infty} g_i(z) \times g_j(y - z)\, dz, \tag{83}$$

$$g_N(\beta_1) = \circledast_{i=1}^{n} [c_i \times f_i(y_i)], \tag{84}$$

$$G_N(\beta_1) = \int_{\check{\beta}_1}^{\hat{\beta}_1} g(\beta_1)\, d\beta_1, \tag{85}$$

where the subscripts i and j in Equation (83) represent two different distributions. Initially, this may be two individual samples. However, this operation must be repeated a total of $N - 1$ times where N represents the total number of samples relevant to $\mathbb{T}4$. For example, if $N = 3$, then the first application of Equation (83) may combine the first two samples, and the second application will then combine the third sample with the output of the first application. Finally, one can use the derived CDF to calculate the probability of passing $\mathbb{T}4$ as shown in Equation (86):

$$P(T_4) = P(\beta_1 < \tau_4). \tag{86}$$

Even if assuming samples and gases are independent, the probabilities of passing each table cannot be considered as such. For example, a gas measurement cannot concurrently fail 𝕋2 whilst passing 𝕋1. Therefore, Equations (73)–(75), and (86) describing the probabilities of passing each table cannot be used directly to determine **L**. Joint probabilities are required, which will differ for **S1–3**. These are discussed here.

S1: Initial Sampling Protocol, (𝕋1–2)

𝕋1–2 are both dependent on Y_1 and must be considered together when determining the probability of each **L**. This is shown in the top plot of Fig. 5-4. $\check{Y}$, $\hat{Y}$, and $\bar{Y}$ represent the minimum, maximum and mean values, respectively, and are highlighted via dashed vertical lines. The ordinate plots the PDF which in this case is a triangular distribution. The relevant limits are then plotted as solid vertical lines labelled as τ_1 and τ_2. Fig. 5-4 is using contrived values purely for demonstrative purposes. To determine the probability of **L1–3** using Equations (65)–(67), Equations (87)–(89) must be used. These assume $\tau_1 \leq \tau_2$, and represent the relevant joint distributions for each **L** which are also shaded in Fig. 5-4 in accordance with the specific example.

$$P(L=1) = \begin{cases} 1 & if\ \hat{Y}_1 < \tau_1, \\ 0 & if\ \tau_1 \leq \check{Y}_1, \\ F_1(\tau_1) - F_1(\check{Y}_1) & otherwise, \end{cases} \tag{87}$$

$$P(L=2) = \begin{cases} 0 & if \begin{vmatrix} \hat{Y}_1 < \tau_1, \\ \tau_2 \leq \check{Y}_1, \end{vmatrix} \\ 1 & if\ \left(\hat{Y}_1 < \tau_2\right) \cap \left(\tau_1 \leq \check{Y}_1\right), \\ F_1(\hat{\gamma}_1) - F_1(\check{\gamma}_1) & otherwise, \end{cases} \tag{88}$$

$$P(L=3) = \begin{cases} 1 & for\ \tau_2 \leq \check{Y}_1, \\ 0 & for\ \hat{Y}_1 < \tau_2, \\ F_1\left(\hat{Y}_1\right) - F_1(\tau_2) & otherwise, \end{cases} \tag{89}$$

where γ are conditional limits based the requirements outlined in Equation (66). For example, the lower limit, $\check{\gamma}_1$, here would be equal to $max(\tau_1, \check{Y}_1)$. A secondary term, Υ, will be used where there are two conditional limits in an equation. The definitions for each γ and Υ are being omitted for brevity but are explained in Annex B.

S2: Periodic Sampling Protocol, (𝕋1– 3)

𝕋1–3 similarly cannot be considered independent as they depend on the most recent sample, Y_1. A more generalised form of Equation (77), shown in Equation (90), can be used, where $P(\star)$ represents the entire attainable probability space.

$$P(\star) = \underbrace{\underbrace{\int_{\check{Y}_1}^{\hat{Y}_1} f_1(y_1)}_{\tau_1,\tau_2} \times \int_{\check{Y}_2}^{\hat{Y}_2} f_2(y_2)\, dy_2 dy_1}_{\tau_3}. \tag{90}$$

The integration across Y_1 relates to 𝕋1–2, and the double integral that also integrates across Y_2 relates to 𝕋3. This is equivalent to Equation (77) but without the assumption that 𝕋3 must pass. The objective is to quantify the relevant regions that satisfy the conditions representing the different possible outcomes regarding 𝕋1–3. However, there are several permutations to consider as the relative positions of the integral limits depend on the relevant table limits, τ, and the measured gas values, Y.

The bottom plot of Fig. 5-4 illustrates the probability space where the most recent sample, Y_1, is shown on the abscissa and the previous sample, Y_2, is shown on the ordinate. In this example, triangular distributions are again assumed but only their mid-points and limits are shown as dashed lines. The two vertical solid lines represent the limits for 𝕋1–2, and the diagonal solid line is the limit for 𝕋3. To 'pass' the tables, the Y_1 must be to the left of the three respective limits with an additional stipulation that to pass 𝕋3, Y_2 must also be above the drawn 𝕋3 limit. There are then three regions outlined, representing **L1–3**. While specifics vary by case, the general process remains identifying and integrating relevant regions. Equations (91), (92), and (93) provide generalised forms of the integrals for **L1–3**.

$$P(L=1) = \begin{cases} 1 & if\ \left(\hat{Y}_1 < \tau_1\right) \cap \left(\hat{Y}_1 - \check{Y}_2 < \tau_3\right), \\ 0 & if\ \left(\tau_1 \le \check{Y}_1\right) \cup \left(\tau_3 \le \check{Y}_1 - \hat{Y}_2\right), \\ \underbrace{\int_{\check{Y}_1}^{\hat{Y}_{1,L1}} f(y_1) \times \int_{\check{Y}_{2,L1}}^{\hat{Y}_2} f(y_2) dy_2 dy_1}_{P\left(T_{1,g} \cap T_{3,g}\right)} & otherwise, \end{cases} \tag{91}$$

$$P(L=2) \tag{92}$$

$$= \begin{cases} 1 & if \begin{vmatrix} (\tau_1 \leq \check{Y}_1) \cap (\hat{Y}_1 < \tau_2), \\ (\hat{Y}_1 < \tau_1) \cap (\tau_3 \leq \check{Y}_1 - \hat{Y}_2), \end{vmatrix} \\ 0 & if \begin{vmatrix} (\tau_2 \leq \check{Y}_1), \\ (\hat{Y}_1 < \tau_1) \cap (\hat{Y}_1 - \check{Y}_2 < \tau_3), \\ (\tau_1 \leq \check{Y}_1) \cap (\tau_3 \leq \check{Y}_1 - \hat{Y}_2), \end{vmatrix} \\ \underbrace{\int_{\check{Y}_1}^{\hat{Y}_{1,L2}} f(y_1) \times \int_{\check{Y}_2}^{\hat{Y}_{2,L2}} f(y_2)\, dy_2 dy_1}_{P(T_{1,g} \cap \neg T_{3,g})} + \underbrace{\int_{\check{Y}_{1,L2}}^{\hat{Y}_{1,L2}} f(y_1)\, dy_1}_{P(\neg T_{1,g} \cap T_{2,g})} & otherwise, \end{cases}$$

$$P(L=3) = \begin{cases} 1 & if\ \tau_2 \leq \check{Y}_1, \\ 0 & if\ \hat{Y}_1 < \tau_2, \\ \underbrace{\int_{\tau_2}^{\hat{Y}_1} f(y_1)\, dy_1}_{P(\neg T_{2,g})} & if\ \check{Y}_1 < \tau_2 \leq \hat{Y}_1, \end{cases} \tag{93}$$

where γ and Υ are conditional limits based on the requirements outlined in Equations (67)–(69). As discussed, Annex B has more details on γ and Υ. If a limit's overriding condition changes part-way through an integral, it should be split into an integration by parts. Equations (62)–(64) would derive **Ł1–3** based on these outputs for **L1–3**.

S3: Periodic Sampling Protocol, (𝕋1–4)

The final scenario to consider is where all tables are required. As with the other cases, the marginal form isolating solely 𝕋4 is not very useful and must be partitioned to allow interrogation of samples Y_1 and Y_2 for testing 𝕋1–3. As mentioned in [47], it is only the first two samples that need to be treated separately, and the remaining samples can be combined into a single marginal distribution, here termed as ℕ, as shown in Equation (94). This is done with the same method as explained for Equations (83)–(85). Please note that ℕ is distinct to N as it is all **remaining** samples sans the most recent two, i.e., $\mathbb{N} = N - 2$. The equivalent to Equation (90) that showed the probability space for 𝕋1–3, $P(\star)$, is shown in Equation (95) for 𝕋1–4.

$$G_N(\beta_1) = \int_{\check{Y}_1}^{\hat{Y}_1} g_1(y_1) \times \int_{\check{Y}_2}^{\hat{Y}_2} g_2(y_2) \times \int_{\check{Y}_{\mathbb{N}}}^{\hat{Y}_{\mathbb{N}}} g_{\mathbb{N}}(y_{\mathbb{N}})\, dy_{\mathbb{N}}\, dy_2\, dy_1, \tag{94}$$

$$P(\star) = \underbrace{\underbrace{\int_{\check{Y}_1}^{\hat{Y}_1} f_1(y_1)}_{\tau_1,\tau_2} \times \underbrace{\int_{\check{Y}_2}^{\hat{Y}_2} f_2(y_2)}_{\tau_3} \times \underbrace{\int_{\check{Y}_{\mathbb{N}}}^{\hat{Y}_{\mathbb{N}}} g_N \left[\sum_{i \in \{1,2,\mathbb{N}\}} (c_i \times y_i) \right] dy_{\mathbb{N}}}_{\tau_4}}_{\tau_4} \; dy_2 \; dy_1. \tag{95}$$

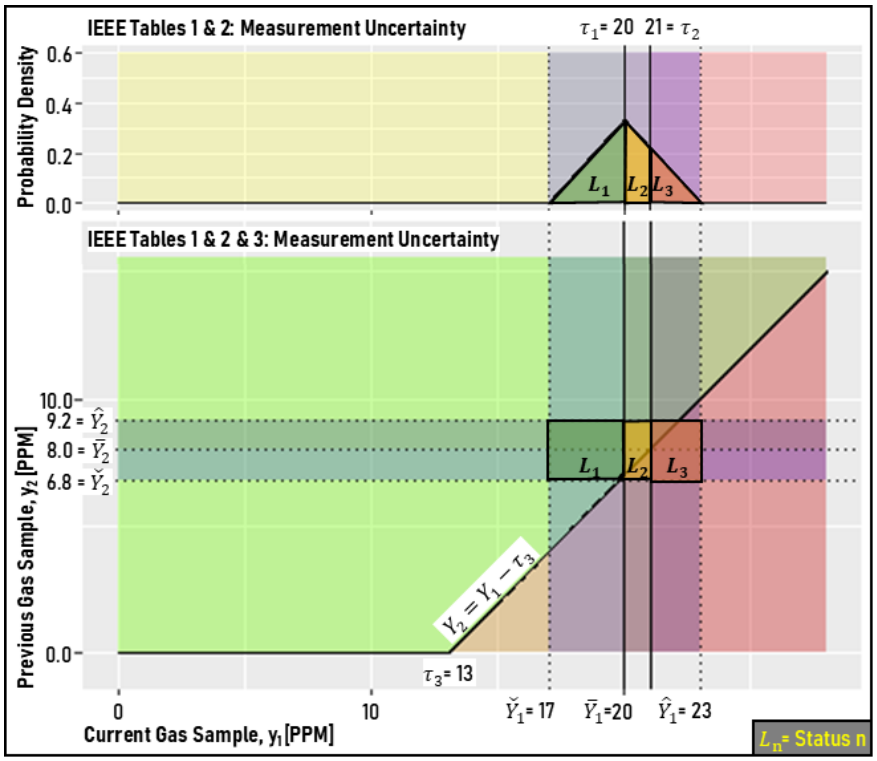


Fig. 5-4: Graphical Determination of Uncertainty for IEEE Methodology's Tables 1 and 2 (Top) and Tables 1 through 3 (Bottom)

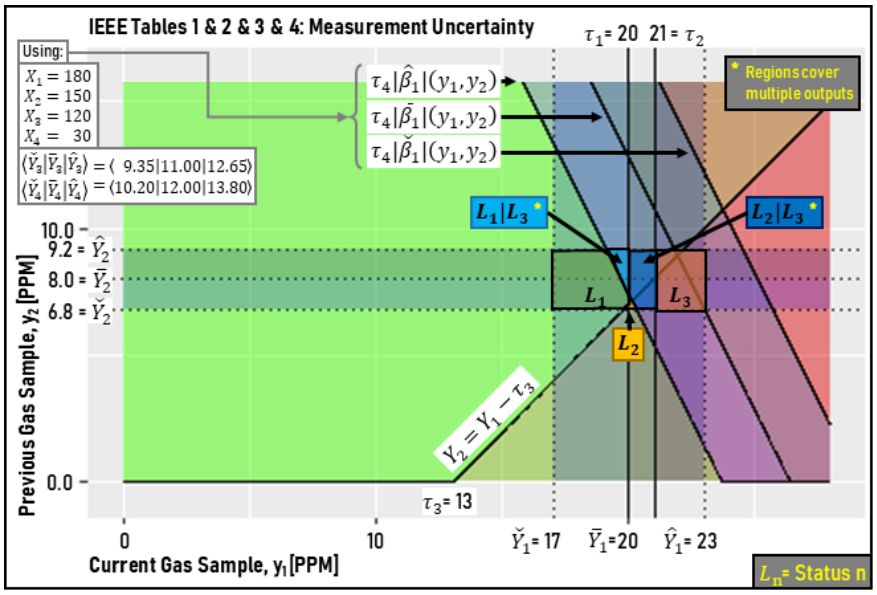


Fig. 5-5: Graphical Determination of Uncertainty for IEEE Methodology's Tables 1 through 4

To aid with intuition, Fig. 5-5 illustrates the problem via a variation of Fig. 5-4 where there is now an additional bounded diagonal zone. This newly added zone is based on a form of Equation (96) where at any given point, Y_1 and Y_2 are assumed fixed, but the combined marginal distribution of $Y_{\mathbb{N}}$ is transposed onto the τ_4 limit. In other words, where limits for $\mathbb{T}1$–3 represented a binary divide between passing or failing, $\mathbb{T}4$'s limit is represented as a distribution given the Y_1 and Y_2. Equations (81) and (82) can be adjusted to calculate the extents of β_1 given values of Y_1 and Y_2.

$$P(T_4|(y_2|y_1)) = P(\beta_1|(y_2|y_1) < \tau_4). \tag{96}$$

There are therefore three parallel lines in Fig. 5-5 where the bottom-left line and top-right lines represent the limits below and above which $\mathbb{T}4$ cannot fail or pass, respectively. The middle line then represents where passing and failing $\mathbb{T}4$ is considered equally likely. This means that in Fig. 5-5, there are two regions highlighted in blue that have a mixed probability of different **L** outcomes whereas other zones have definitive outcomes. Again, this is because $Y_{\mathbb{N}}$ is being treated as a distribution at any given point whereas Y_1 and Y_2 are treated as fixed values y_1 and y_2. Fig. 5-5 is representative of the scenario where $x_2 > \bar{x}_N$; were it that $x_2 = \bar{x}_N$ or $x_2 < \bar{x}_N$ then the slope of $\mathbb{T}4$'s zone would be vertical or positive, respectively. This means that an increase in Y_2 can either increase, decrease, or have no impact on β_1, depending on whether the sampling time, x_2, of Y_2 is greater than, less than, or equal to the average sampling time, respectively. In contrast, $x_{\mathbb{N}} < \bar{x}_{\mathrm{N}}$ can be assumed.

This intermediate zone is represented by a distribution shape that may differ from the zones representing Y_1 and Y_2, depending on the convolution described in Equation (83). Specifically, if triangular distributions are assumed, the distribution representing τ_4 in Fig. 5-5 will be triangular only if $\mathbb{N} = 1$. If a $\mathcal{N}$ distribution is assumed, then so will τ_4's distribution. The transposition of part of β_1 onto τ_4 allows projection onto a two-dimensional plot. Fig. 5-5 uses an offset of $X_4 = 30$ days to have an interesting intersection but the correct usage would always have the oldest sample at $X_N = 0$.

As there are more permutations to consider, the equations become too unwieldly to include all explicitly. The equations describing **L1–3** are split up such that Equations (97), (98), and (99) first isolate the most relevant region in the probability space for **L1–3**, respectively. Then, it is for reader to derive the $\rho(L = i)$ terms by using the generic integral, $P(\star)$, from Equation (95) and Annex B for the appropriate limits.

$$P(L=1)=\begin{cases}1 & if\ \left(\hat{Y}_1<\tau_1\right)\cap\left(\hat{Y}_1-\check{Y}_2<\tau_3\right)\cap\left(\hat{\beta}_1<\tau_4\right),\\ 0 & if\ \left(\tau_1\le\check{Y}_1\right)\cup\left(\tau_3\le\check{Y}_1-\hat{Y}_2\right)\cup\left(\tau_4\le\check{\beta}_1\right),\\ \rho(L=1) & otherwise,\end{cases} \tag{97}$$

$$P(L=2)=\begin{cases}0 & if\ \begin{vmatrix}\left(\tau_1\le\check{Y}_1\right)\cap\left(\hat{Y}_1<\tau_2\right),\\ \left(\hat{Y}_1<\tau_1\right)\cap\left(\tau_3\le\check{Y}_1-\hat{Y}_2\right),\end{vmatrix}\\ 1 & if\ \begin{vmatrix}\left(\tau_1\le\check{Y}_1\right)\cap\left(\hat{Y}_1<\tau_2\right)\cap\left(\tau_3\le\check{Y}_1-\hat{Y}_2\right)\cap\left(\hat{\beta}_1<\tau_4\right),\\ \left(\hat{Y}_1<\tau_2\right)\cap\left(\tau_3\le\hat{Y}_1-\check{Y}_2\right)\cap\left(\hat{\beta}_1<\tau_4\right),\end{vmatrix}\\ \rho(L=2) & otherwise,\end{cases} \tag{98}$$

$$P(L=3)=\begin{cases}0 & if\ \begin{vmatrix}\left(\hat{Y}_1<\tau_1\right)\cap\left(\hat{Y}_1-\check{Y}_2<\tau_3\right),\\ \left(\tau_2\le\check{Y}_1\right),\\ \left(\tau_4\le\check{\beta}_1\right)\end{vmatrix}\\ 1 & if\ \begin{vmatrix}\left(\tau_2\le\check{Y}_1\right),\\ \left(\tau_4\le\check{\beta}_1\right),\end{vmatrix}\\ \rho(L=3) & otherwise.\end{cases} \tag{99}$$

Equations (62)–(64) would derive **Ł1–3** using **L1–3**.

5.2.2. Attempted Algebraic Solution

This Sub-Section considers algebraic approaches for calculating **Ł** using symmetric triangular, Δ, or $\mathcal{N}$ distributions to conclude that it is unlikely a simple one exists.

DGA Sample Distributions

Triangular DGA Sample Distribution

The publication, [47], used a Δ distribution. However, evaluation is laborious due to the multiple equations defining the distribution. Equations (100) and (101) describe the PDF, $f_i^{\Delta}(y_i)$, and CDF, $F_i^{\Delta}(y_i)$, for a triangular distribution, respectively.

$$f_i^{\Delta}(y_i)=\begin{cases}0 & if\ y_i<\check{Y}_i,\\ 4\left(y_i-\check{Y}_i\right)(W_i)^{-1} & if\ \check{Y}_i\le y_i<\bar{Y}_i,\\ 2(W_i)^{-1} & if\ y_i=\bar{Y}_i,\\ 1-4\left(\hat{Y}_i-y_i\right)(W_i)^{-1} & if\ \bar{Y}_i<y_i\le\hat{Y}_i,\\ 1 & if\ \hat{Y}_i<y_i,\end{cases} \tag{100}$$

$$F_i^{\Delta}(y_i)=\begin{cases}0 & if\ y_i<\check{Y}_i,\\ 2\left(y_i-\check{Y}_i\right)^2(W_i)^{-1} & if\ \check{Y}_i<y_i\le\bar{Y}_i,\\ 1-2\left(\hat{Y}_i-y_i\right)^2(W_i)^{-1} & if\ \bar{Y}_i<y_i<\hat{Y}_i,\\ 1 & if\ \hat{Y}_i\le y_i,\end{cases} \tag{101}$$

where the superscripted Δ denotes it is describing a Δ distribution and is **not** intended as an exponent. W represents the range between the minimum and maximum potential values for a given sample, Y. This assumes that the *IEC specification*'s ±15% is natively expressing the relevant range for the triangular distribution.

Normal DGA Sample Distribution

The $\mathcal{N}$ distribution is the other candidate distribution, where Equations (102) and (103) describe $f_i^{\mathcal{N}}(y_i)$ and $F_i^{\mathcal{N}}(y_i)$, respectively. $F^{\mathcal{N}}$is presented as such in Equation (103) as it does not have a general closed-form solution. However, its close relation to the *Error Function* [107] means that it is *Analytic* [108]. In practice, its ubiquity has led to many calculators / software packages supporting its evaluation.

$$f_i^{\mathcal{N}}(y_i) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i-\bar{Y}_i}{\sigma_i}\right)^2}, \tag{102}$$

$$F_i^{\mathcal{N}}(y_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\left(\frac{y_i-\bar{Y}_i}{\sigma_i}\right)} e^{-t^2} dt. \tag{103}$$

where the superscripted $\mathcal{N}$ denotes it is describing a $\mathcal{N}$ distribution and is **not** intended as an exponent. It is assumed that the *IEC specification*'s ±15% represents a *Confidence Interval* derived from a *Confidence Level* of 95%, equivalent to a *Coverage Factor* of $k = 1.96$. Equation (104) defines the assumed standard deviation, σ, for the $\mathcal{N}$ distribution.

$$\sigma_i = \alpha \times \bar{Y}_i / k, \tag{104}$$

where α is equal to 0.15. Note that this interpretation is not equivalent to that of the triangular distribution, rather, Equation (105) provides this approximate equivalence.

$$\sigma_i^{\Delta} \sim \sqrt{W_i^2/24}. \tag{105}$$

Marginal Distributions of Tables

Absolute Gas Values, (𝕋𝟙–𝟚)

Equations (73) and (74) describing the probabilities of passing 𝕋𝟙–𝟚 can be solved directly using either Equation (101) or (103), depending on the desired distribution.

Delta in subsequent Absolute Gas Values, (𝕋𝟛)

Equation (77) describing the probability of passing 𝕋𝟛 can be handled differently depending on the desired distribution. For $\mathcal{N}$ distributions, the distribution shape does not change when subtracting other $\mathcal{N}$ distributions of differing parameters, i.e., Y_{1-2}

can also be assumed a $\mathcal{N}$ distribution with a mean and standard deviation as defined by Equations (106) and (107), respectively. Therefore, Equations (102) and (103) can be trivially adjusted to represent this new distribution to solve Equation (75), giving the probability of passing $\mathbb{T}\mathbb{3}$.

$$\bar{Y}_{1-2} = \bar{Y}_1 - \bar{Y}_2, \tag{106}$$

$$\sigma_{1-2} = \sqrt{({\sigma_1}^2 + {\sigma_2}^2)}. \tag{107}$$

However, for triangular distributions, although the expected mean, minimum, and maximum values for the delta are calculable via Equations (106), (108), and (109), respectively, the shape of this marginal distribution is not triangular. Therefore, they cannot be used directly in Equations (100) and (101) to solve Equation (75).

$$\check{Y}_{1-2} = \check{Y}_1 - \hat{Y}_2, \tag{108}$$

$$\hat{Y}_{1-2} = \hat{Y}_1 - \check{Y}_2. \tag{109}$$

To determine the marginal distribution shape, Y_1 and $-Y_2$ can be convolved. However, the single integral method shown in Equation (77) is pursued. Fig. 5-6 illustrates the problem. The diagonal line demarcates the $y_2 > y_1 - \tau_3$ threshold, dividing where the potential combinations of sample values can pass $\mathbb{T}\mathbb{3}$ from where they cannot. These generic regions are shaded in green and red, respectively. The problem space can be constrained further to include only the region where both y_1 and y_2 are obtainable. This space is divided into four zones, labelled **A–D** in Fig. 5-6. As triangular distributions must be split into their lower and upper halves, each zone represents one of the four potential combinations. Depending on the gas values and τ_3, the diagonal line can lie across between zero and three of these regions. In the example shown in Fig. 5-6, the relevant segments along the Y_1 domain are labelled as α, β, γ, δ, and ω.

The segment ω is the simplest to evaluate as it is where either an unobtainable value for Y_1 or Y_2 is required and thus can be equated to a probability of zero. The segment α is also simple as it is where any value for Y_2 for a given y_1 results in passing $\mathbb{T}\mathbb{3}$. Therefore, this segment reverts to the equivalent of F_1 obtainable via Equation (101). The remaining three segments, β, γ, and δ, would in this example correspond to zones **B**, **C**, and **D**, respectively. As explained for segments α and ω, only when the lower bound is dictated by τ_3 are the solutions non-trivial. Therefore, Equations (110)–(113) describing each of the four zones, respectively, focus specifically on this scenario. The final probability will be the sum of all relevant integrals.

$$P_A(Y_{1-2} < \tau_3) = \left[\frac{2(y_1 - \hat{Y}_2 - \tau_3)^3(3y_1 + \hat{Y}_2 - 4\check{Y}_1 + \tau_3)}{3(\hat{Y}_1 - \check{Y}_1)^2(\hat{Y}_2 - \check{Y}_2)^2}\right]_{\check{y}_1}^{\hat{y}_1}, \tag{110}$$

$$P_B(Y_{1-2} < \tau_3) = \left[-\frac{2(y_1 - \check{Y}_2 - \tau_3)^4}{(\hat{Y}_1 - \check{Y}_1)^2(\hat{Y}_2 - \check{Y}_2)^2} - \frac{8(\check{Y}_2 - \check{Y}_1 + \tau_3)(y_1 - \check{Y}_2 - \tau_3)^3}{3(\hat{Y}_1 - \check{Y}_1)^2(\hat{Y}_2 - \check{Y}_2)^2} + \frac{2(y_1 - \check{Y}_2 - \tau_3)^2 + 4y_1(\check{Y}_2 - \check{Y}_1 + \tau_3)}{(\hat{Y}_1 - \check{Y}_1)^2}\right]_{\check{y}_1}^{\hat{y}_1}, \tag{111}$$

$$P_C(Y_{1-2} < \tau_3) = \left[\frac{2(y_1 - \check{Y}_2 - \tau_3)^4}{(\hat{Y}_1 - \check{Y}_1)^2(\hat{Y}_2 - \check{Y}_2)^2} + \frac{8(\check{Y}_2 - \hat{Y}_1 + \tau_3)(y_1 - \check{Y}_2 - \tau_3)^3}{3(\hat{Y}_1 - \check{Y}_1)^2(\hat{Y}_2 - \check{Y}_2)^2} + \frac{2(y_1 - \check{Y}_2 - \tau_3)^2 + 4y_1(\check{Y}_2 - \check{Y}_1 + \tau_3)}{(\hat{Y}_1 - \check{Y}_1)^2}\right]_{\check{y}_1}^{\hat{y}_1}, \tag{112}$$

$$P_D(Y_{1-2} < \tau_3) = \left[-\frac{2(y_1 - \hat{Y}_2 - \tau_3)^3(3y_1 + \hat{Y}_2 - 4\hat{Y}_1 + \tau_3)}{3(\hat{Y}_1 - \check{Y}_1)^2(\hat{Y}_2 - \check{Y}_2)^2}\right]_{\check{y}_1}^{\hat{y}_1}, \tag{113}$$

where $\check{y}_1$ and $\hat{y}_1$ represent the limits for a scenario. Equations (91)–(93) for **L1–3** are solvable algebraically via the use of Equations (110)–(113). Although, the process would be cumbersome with many segmented regions to integrate.

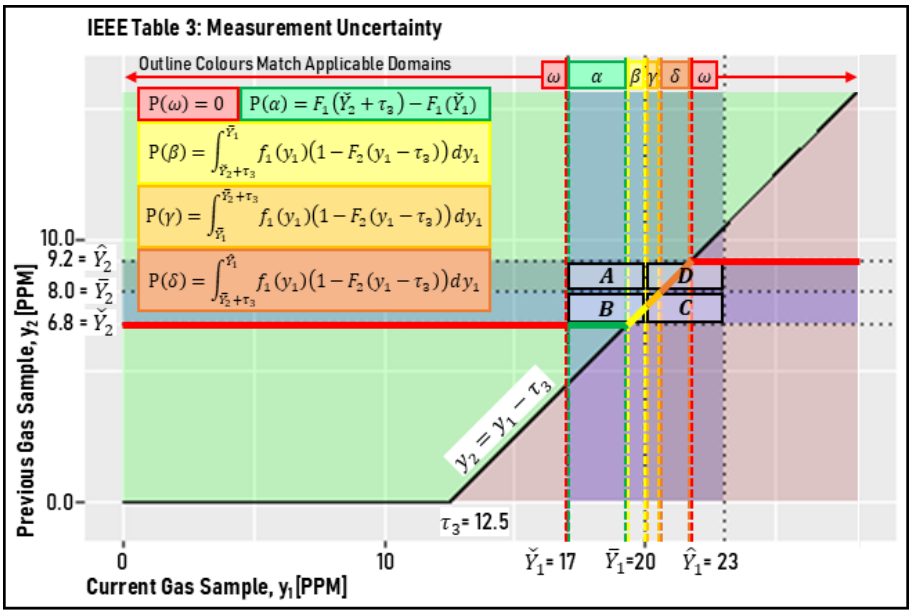


Fig. 5-6: Graphical Representation of Calculating Uncertainty for IEEE Methodology's Tables 1 through 3

Average Gassing Rate, (𝕋4)

Equation (86) describing the probability of passing 𝕋4 can also be handled differently depending on the desired distribution. Following a similar logic, $\mathcal{N}$ distributions combined via convolution result in a $\mathcal{N}$ distribution. Therefore, Equations (102) and

(103) can be adjusted to represent the new distribution in Equation (84) to solve Equation (86), giving the probability of passing 𝕋4. A simple modification to Equation (78) utilising Equation (107), as shown in Equation (114), allows for the calculation of the *Standard Error of Mean*, $SEM_{\beta 1}$, which here is assumed the *Measurement Uncertainty*.

$$SEM_{\beta 1} = \sqrt{\sum_{i=1}^{n} (c_i \times \sigma_i)^2}, \tag{114}$$

where σ_i represents the standard deviation of the $\mathcal{N}$ distribution representing sample Y_i. This can be used with Equation (78) to describe the $\mathcal{N}$ distribution's mean. An alternative is using the R package "Metafor" [109] and its "RMA" function. This function can calculate Equation (114) when using its "unweighted" variant of the "Fixed Effect" model if assuming $\mathcal{N}$ distributions. Although its algorithm to derive this value was not investigated, its answer is nevertheless the same. It is assumed sufficient samples to ignore the *Effective Degrees of Freedom*.

Unfortunately, this simplification is not applicable for the case of assuming triangular distributions and instead Equations (84) and (85) are needed where, again, the summation must be done via convolution as per Equation (83). There does not appear to be any elegant short-form expression for the outputs of said convolutions for triangular distributions, especially as the number of iterations increase. Indicatively, the number of segments describing the resulting PDF will be $2 + 2^n$. It is therefore argued that a strict algebraic solution assuming triangular distributions is impractical as the overall length of the equations required grows prohibitively cumbersome.

Joint Distributions of Tables

S1: Initial Sampling Protocol, (𝕋1–2)

𝕋1–2 and trivially combined for both distributions.

S2: Periodic Sampling Protocol, (𝕋1–3)

For a triangular distribution, the overall approach would remain the same as for a marginal distribution shown in Equations (110)–(113)., but with more segmentations to represent a given **L**, described in Equations (91)–(93). However, this algebraic solution is arguably already impractical given the number of integrations required.

Despite the relative ease of deriving the marginal distributions for the case of $\mathcal{N}$ distributions, the joint distributions become more involved. If framed as a bivariate normal distribution, it can be readily evaluated with relevant software packages. This can be done primarily because Y_1 and Y_2 are considered independent, so a bivariate representing Y_1 and $Y_1 - Y_2$ can be used where the correlation is known as an increase in Y_1 leads to an increase in $Y_1 - Y_2$. The bivariate can be rescaled to a more standard distribution that has pre-tabulated values such as in [110] that was used in [111]. This shown in Equation (115) and is a slight modification of the method explained in [111], where the changes were to isolate the region representing **L1**. The **L3** is the same as the case for **S1** and so would be trivially solvable using Equation (103). **L2** can most simply be obtained by equating it to the residual not covered by **L1** or **L3**.

$$P(L=1) = \int_{-\infty}^{h}\int_{-\infty}^{k} \frac{e^{-\left[\frac{1}{2\sqrt{1-r^2}}\left(y_2^2-(2\times r\times y_2\times y_1)+y_1^2\right)\right]}}{2\pi\sqrt{1-r^2}}\, dy_2\, dy_1, \tag{115}$$

$$h = (\tau_1 - \bar{Y}_1)/\sigma_1, \tag{116}$$

$$k = \left(\tau_3 - (\bar{Y}_1 - \bar{Y}_2)\right)/\sqrt{{\sigma_1}^2 + {\sigma_2}^2}, \tag{117}$$

$$r = \sigma_1/\sqrt{{\sigma_1}^2 + {\sigma_2}^2}. \tag{118}$$

where h, k, and r are given by Equations (116), (117), and (118), respectively.

S3: Periodic Sampling Protocol, (𝕋1–4)

As with the marginal distribution, solving the joint distribution algebraically with triangular distributions seem impractical. Indicatively, a triangular distribution repeatedly convolved will require an integration by parts segmented $2 + 2^{N-1}$ times, where N here ranges between 3–6. Similarly, the $\mathcal{N}$ distributions also seem impractical to solve analytically given the complexity. Although $Y_{\mathbb{N}}$ can be derived, the overall scenario represents a truncated trivariate normal distribution with complicating interdependencies. The joint distribution of 𝕋3 had a simple linear relationship between Y_1 and $Y_1 - Y_2$. A similarly simple relationship between Y_1 and β_1, and of Y_2 and β_1 can be derived. However, the relationship between $Y_1 - Y_2$ and β_1 is different in being a non-linear relationship as an increased delta could be due to either an increased Y_1 or a decreased Y_2, each impacted by β_1 differently. Searching for an equivalent to Equation (103) may be possible but was considered out of scope for this thesis.

5.2.3. Proposed Numerically Estimated Methodologies

Thus, neither distribution present a practical algebraic solution. Instead, numerical estimation is pursued, considering two viable approaches: a *Monte Carlo Method* (MCM), and numerical integration. Both provide outputs alongside an estimate of its accuracy, which can be adjusted at the cost of processing time to reach a desired tolerance. As the propagation of *Measurement Uncertainty* is a means to convey potential outcomes **not** explicitly measured, there is no ground truth. Therefore, the focus is on comparing the results between the two approaches, with any discrepancies being discussed. These two reference points can also validate the assertions made with the above equations.

Input Data: M1

A contrived dataset, labelled **M1**, is used, consisting of a set of 6 measurements for one gas. Given that independence between gases is assumed, the derivation of **Ł** is trivial and skipped for now. **M1** has associated with it two variant cases, **CA** and **CB**, representing different limits for $\mathbb{T}\mathbb{1}$–$\mathbb{4}$. These values are shown in Table 5-1. Three distributions will be compared, a $\mathcal{N}$ distribution, a symmetric triangular (Δ) distribution, and a scaled $\mathcal{N}$ distribution made to match the Δ distribution, $\mathcal{N}^{\Delta}$. These were explained for Equations (100)–(105). For these distributions, the samples can either be each processed individually or have samples 3–6 combined into an equivalent sample $\mathbb{N}$, as explained for Equation (94). These processing techniques will be termed, N and $\mathbb{N}$, respectively. For these variations, *Protocols* **S1**–**3** will be considered.

*Table 5-1: Case Study Data **M1** with Limit Variants **CA** and **CB***

Index	***Date**	$\bar{Y}_i$	σ_i	σ_i^{Δ}	W_i	c_i	**Protocol**	**Table**	**Case**	
i	Days	PPM	PPM	PPM	PPM	x10^{-4}	Sn	**Limits**	CA	CB
1	720	20.00	1.53	1.22	3.00	9.92	1+2+3	τ_1 [PPM]	20.0	19.0
2	576	^8.00	^0.77	^0.61	^1.50	5.95	2+3	τ_2 [PPM]	21.0	20.0
3	432	11.00	0.84	0.67	1.65	1.98	3	τ_3 [ΔPPM]	12.5	11.5
4	288	12.00	0.92	0.73	1.80	-1.98	3	τ_4 [ΔPPM/Yr]	4.0	3.0
5	144	10.00	0.77	0.61	1.50	-5.95	3			
6	0	^9.00	^0.17	^0.61	^1.50	9.92	3			

***:** Calculated as days since oldest sample was taken.
^: Increased Uncertainty as nearing LoD as per Eq. (56). ($U_{i=1}$ = 18.75%, $U_{i=6}$ = 16.67%).

Processing Methodology

<u>Numerical Integration</u>

Using R, $f^{\mathcal{N}}$ and $F^{\mathcal{N}}$ were calculated via the default functions “dnorm” and “pnorm”, respectively. For f^{Δ} and F^{Δ}, the functions “dtriangle” and “ptriangle” from the package

"triangle" [112] was used. An example is shown in Equation (119) showing the probability of passing $\mathbb{T}\mathbb{1}$ assuming $\mathcal{N}$.

$$P(Y_1 < \tau_1) \approx pnorm(\tau_1, \bar{Y}_1, \sigma_1). \tag{119}$$

To estimate an arbitrary numerical integration, the package "pracma" [113] was used with its functions: "integral", "integral2", and "integral3", each referring to the number of dimensions it is integrating across. The first function's implementation is based on [114] and the other two based on [115], and they are all based on the well-established Gauss-Konrod [116] method for adaptive numerical integration. A function defining the integral is required, along with any relevant limits, with optional arguments for error tolerances. Where possible, the number of dimensions being integrated was minimised. This process was explained for Equations (76) and (77). An example is shown in Equation (120) showing the probability of passing $\mathbb{T}\mathbb{3}$ assuming $\mathcal{N}$.

$$P(Y_{1-2} < \tau_3) \approx integral\left[\prod \begin{vmatrix} dnorm(y_1, \bar{Y}_1, \sigma_1) \\ 1 - pnorm(y_1 - \tau_3, \bar{Y}_2, \sigma_2) \end{vmatrix}_{\check{Y}_1}^{\hat{Y}_1}\right]. \tag{120}$$

Another example, in Equation (121), shows the probability of passing $\mathbb{T}\mathbb{4}$ assuming $\mathcal{N}$. For this approach to work, the combined $Y_{\mathbb{N}}$ must be a valid simplification. This will be demonstrated shortly via the MCM.

$$P(\beta_1 < \tau_4) \approx integral2\left[\prod \begin{vmatrix} dnorm(y_1, \bar{Y}_1, \sigma_1) \\ dnorm(y_2, \bar{Y}_2, \sigma_2) \\ pnorm([y_1 \times c_1] + [y_2 \times c_2] - \tau_4, -\bar{Y}_{\mathbb{N}}, \sigma_{\mathbb{N}}) \end{vmatrix}_{\check{Y}_1, \check{Y}_2}^{\hat{Y}_1, \hat{Y}_2}\right]. \tag{121}$$

Despite the apparent simplicity of this approach and fast processing time, there are issues regarding numerical stability that can lead to unexpected results. Limits must be handled carefully to avoid regions of extremely low probabilities to mitigate this issue. Furthermore, the underlying integral equations must still be known to estimate the outputs. As demonstrated later, this becomes a limiting factor for this approach.

Monte Carlo Method

To simulate drawing from given distributions, the default function "rnorm" and the "rtriangle" function from the "triangle" [112] package in R were used. A given number of randomly selected samples are first drawn, and they are then tested against the applicable requirements for either $\mathbb{T}\mathbb{1}$–$\mathbb{4}$ or **L1–3**. The count of samples meeting a given set of criteria can be divided by the total number of samples to give an unbiased

approximate estimate of the probability of achieving said criteria. As the number of 'trials' increases, the approximation converges upon the true value for the probability.

As each trial represents an instance of what is possible, no further calculations are needed to determine joint probabilities, avoiding some complications associated with algebraic solutions. An indicative overview of the implementation is shown in Fig. 5-7. Fig. 5-7 also references a *Diagnosis* step, but this will only be included in Sub-Section 5.2.4 onwards. All relevant outputs required to construct the *DGA Status* for **S1**–**3** are recorded during the same trial to maximise consistency across results.

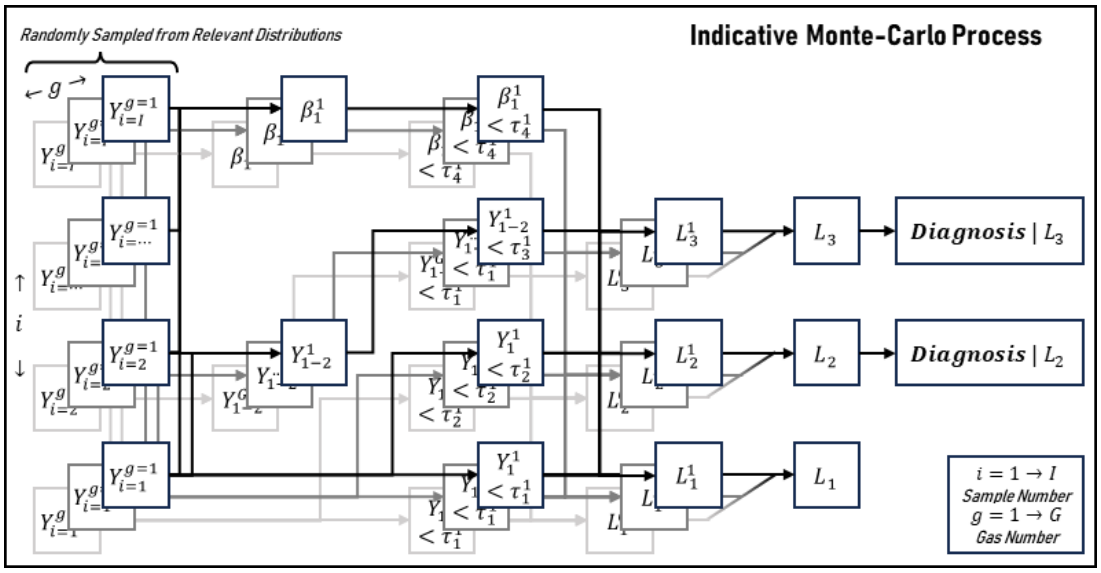


Fig. 5-7: Monte-Carlo Simulation Methodology

To represent a given gas sample's distribution, 10^7 samples were randomly drawn for **M1**. If accommodating that the *Confidence Interval* of passing a table, and of being a given **L**, can be approximated in accordance with the binomial proportion *Confidence Interval*, then there are various equations available to estimate it. Although the Wald interval shown in Equation (122) is traditionally the most typical method, [117] suggests instead the use of the Agresti-Coull interval [118]. This will be used with the inclusion of the slight modification suggested in [117] as shown in Equation (123).

$$W_P = 2k_{1.96}\sqrt{\frac{P(1-P)}{n}}, \tag{122}$$

$$W_P = 2k_2\sqrt{\frac{\left[\frac{2X+{k_2}^2}{2\left(n+{k_2}^2\right)}\right]\times\left(1-\left[\frac{2X+{k_2}^2}{2\left(n+{k_2}^2\right)}\right]\right)}{n+{k_2}^2}}, \tag{123}$$

where W_P is the combined *Confidence Interval* for the given probability of success, P. This is based on the number of successes, X, and the total number of samples, n. Equation (123) uses $k = 2$ for the equivalent coverage of Equation (122)'s $k = 1.96$. Indicatively, both would peak at ±0.03% where $P = 0.5$ given $n = 10^7$, though they differ nearer the tails. It is stated in [45, Sec. 7] that "a value of $M = 10^6$ can often be expected to deliver a 95 % coverage interval for the output quantity such that this length is correct to one or two significant decimal digits". It further advises that ensuring: $n > 10^4/(1 - P)$, will "provide a reasonable discrete representation of ... the regions near the endpoints..." [45, Sec. 7]. $n = 10^7$ would then accommodate $P \approx 0.999$ whilst still meeting the recommendation. It is argued these values are within the order of magnitude as to be insignificant in the context of validating the methodology. For both approaches, n can be increased to satisfy a given desired tolerance for an applied implementation although at the cost of the runtime. Alternatively, [45, Sec. 7] offers information on a method to select the number of trials adaptively.

Results of Dataset M1

Comparing Numerical Estimation Methods

The results for dataset **M1** are shown in Table 5-2. The results include both limits, **CA** and **CB**, as well as *Protocols*, **S1**–**3**. The results show no practical difference across the board between either method of numerical estimation. Indicatively, on a retail laptop with Ryzen 5000 and 16 GB 3200 MHz RAM, the numerical integration took <1 second whereas the MC took <10 seconds. Both times are considered equivalent for practical use-cases. Although, if considering all gases, the difference would be starker. The reason is that the inclusion of all seven gases concurrently causing a non-linear increase in processing times due to the increased data volume. The specific relationship depends on specifics such as the RAM size comparative to the data volume. Nevertheless, Table 5-2 sufficiently demonstrates the viability and validity of the two numerical estimation methods with their matching outputs.

Combining Samples

The motivation for combining samples is for processing times and scalability. As the number of samples considered increases, the computation time will increase non-linearly. For the numerical integration method, it is impractical to scale much beyond three dimensions. However, for the MCM, this is less impactful unless implementing

the previously mentioned modification to $\mathbb{T}4$ to enforce a minimum duration. In such a case, one could have daily samples spanning 4-months. Indicatively, it takes ~40 seconds to evaluate a single gas via MCM using 10^6 trials (was previously 10^7 trials).

For $\mathcal{N}$, combining all samples except the first two into an equivalent $Y_{\mathbb{N}}$ is demonstrated to have no impact on the outputs as shown in Table 5-2. However, for Δ, Table 5-2 shows a potentially significant discrepancy of over 1% in some cases. This discrepancy is due to the implementation which assumes $\mathcal{N}^{\Delta}$ for $Y_{\mathbb{N}}$ rather than convolving. This assumption can be unreliable when there are only between 2–5 convolutions. However, as the number of samples increases further, the repeated convolutions converge towards a normal distribution, allowing for the assumption to be made with minimal impact. At the lower range of samples, the MC method is already fast enough as to not require the combining of samples.

*Table 5-2: Case Study Results of Data **M1** with Limit Variants **CA** and **CB***

Method:	**Simulation**						**Integration**					
Distribution Shape^:	$\mathcal{N}_N$	$\mathcal{N}_{\mathbb{N}}$	$\mathcal{N}_N^{\Delta}$	$\mathcal{N}_{\mathbb{N}}^{\Delta}$	Δ_N	$\Delta_{\mathbb{N}}$	$\mathcal{N}_N$	$\mathcal{N}_{\mathbb{N}}$	$\mathcal{N}_N^{\Delta}$	$\mathcal{N}_{\mathbb{N}}^{\Delta}$	Δ_N	$\Delta_{\mathbb{N}}$
M1: CA: S1:												
$\boldsymbol{P(L=1)}$	50.03	50.01	50.00	50.02	49.99	50.01	–	50.00	–	50.00	–	50.00
$\boldsymbol{P(L=2)}$	24.30	24.32	29.29	29.27	27.79	27.78	–	24.32	–	29.29	–	27.78
$\boldsymbol{P(L=3)}$	25.67	25.67	20.71	20.71	22.22	22.22	–	25.68	–	20.71	–	22.22
M1: CA: S2:												
$\boldsymbol{P(L=1)}$	47.10	47.09	47.78	47.80	47.73	47.74	–	47.07	–	47.78	–	47.74
$\boldsymbol{P(L=2)}$	27.23	27.24	31.51	31.49	30.06	30.04	–	27.25	–	31.51	–	30.04
$\boldsymbol{P(L=3)}$	25.67	25.67	20.71	20.71	22.22	22.22	–	25.68	–	20.71	–	22.22
M1: CA: S3:												
$\boldsymbol{P(L=1)}$	46.17	46.16	47.36	47.38	47.36	46.80	–	46.14	–	47.36	–	46.81
$\boldsymbol{P(L=2)}$	22.68	22.69	27.55	27.53	26.24	25.17	–	22.70	–	27.55	–	25.17
$\boldsymbol{P(L=3)}$	31.15	31.15	25.09	25.09	26.40	28.02	–	31.16	–	25.09	–	28.03
M1: CB: S1:												
$\boldsymbol{P(L=1)}$	25.66	25.67	20.71	20.81	22.21	22.22	–	25.68	–	20.71	–	22.22
$\boldsymbol{P(L=2)}$	24.32	24.33	29.28	29.28	27.79	27.80	–	24.32	–	29.29	–	27.78
$\boldsymbol{P(L=3)}$	50.02	50.00	50.00	50.01	50.01	49.98	–	50.00	–	50.00	–	50.00
M1: CB: S2:												
$\boldsymbol{P(L=1)}$	23.62	23.64	19.37	19.37	20.77	20.78	–	23.63	–	19.37	–	20.78
$\boldsymbol{P(L=2)}$	26.26	26.36	30.63	30.62	29.23	29.24	–	26.37	–	30.63	–	29.22
$\boldsymbol{P(L=3)}$	50.02	50.00	50.00	50.01	50.01	49.98	–	50.00	–	50.00	–	50.00
M1: CB: S3:												
$\boldsymbol{P(L=1)}$	15.68	15.68	12.11	12.10	13.06	12.78	–	15.68	–	12.10	–	12.77
$\boldsymbol{P(L=2)}$	6.84	6.83	5.88	5.88	5.75	6.72	–	6.84	–	5.89	–	6.71
$\boldsymbol{P(L=3)}$	77.48	77.48	82.01	82.02	81.19	80.51	–	77.48	–	80.02	–	80.52

^: $\mathcal{N}$: Normal Dist., Δ: Triangular Distribution, N: Using All Samples, $\mathbb{N}$: Combining Samples.
Note: Confidence Interval < ±0.03% ($n = 10^7$).

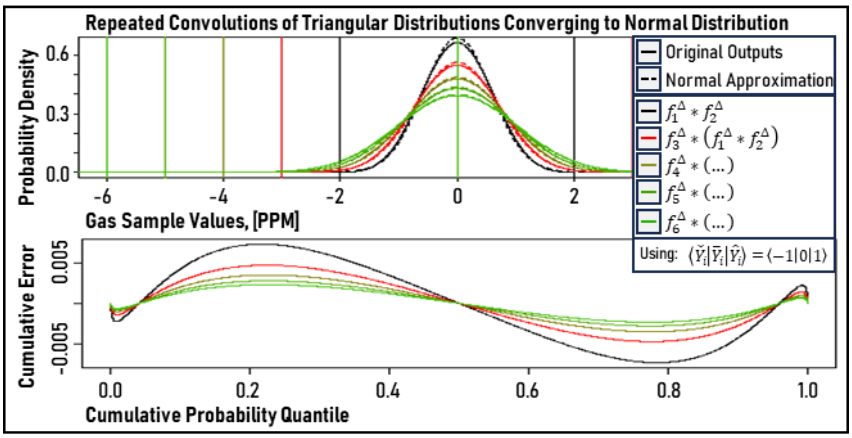


Fig. 5-8: Convolution's Tendency of Convergence Towards Normal Distribution

This assumption considering repeated convolutions to incrementally approximate $\mathcal{N}^{\Delta}$ is generally valid as per the central limit theorem. In practice, the rate of convergence towards $\mathcal{N}^{\Delta}$ will depend on the specific values for each Δ being convolved. Fig. 5-8 demonstrates this convergence by repeatedly convolving a given Δ with itself. The resulting distributions are shown in the top plot where each colour represents a different iteration. The "convolve" function from the "pracma" package [113] in R was used. In the same colour, the dashed line in the top plot shows the approximation of the PDF via $\mathcal{N}^{\Delta}$. The bottom plot then shows the cumulative error across the CDF between this approximation and the original distribution generated via convolved Δ distribution. This bottom plot illustrates the convergence towards $\mathcal{N}^{\Delta}$.

Comparing Distribution Shape

It is challenging to evaluate the comparison of Δ and $\mathcal{N}^{\Delta}$ as it is a borderline tautological matter; being dependent on the validity of the underlying assumption regarding the distribution shape. As there is no inherent ground truth available, they can only be qualitatively compared. One aspect to consider is that $\mathcal{N}$ is much simpler to implement due to the ease at which the combined sample $Y_{\mathbb{N}}$ can be derived. On the other hand, as also mentioned in [119], though Δ is more dubious in a theoretical sense, its intuitiveness due to its bound nature can make it practically appealing. Another use-case is if the *Coverage Factor* of a given interval is not known, but an emphasis on the mid-point is desired as corroborated in Appendix A in [85].

However, given how convolving alters the triangular distributions shape, it is argued that its primary appeal of intuitiveness is somewhat lost as the theoretical bounds are not very representative of the actual resulting distribution. For example, in Fig. 5-8, the

vertical lines demarcate the minimum and maximum potential values after each convolution matched according to colour. This is easily derived by summing either the minimums or maximums, respectively. However, it begins to grossly overestimate the range on a practical level even if it can be argued it may be theoretically possible to reach these extremes, it is at vanishingly small probabilities.

If comparing the CDFs of each distribution shape, the Δ will be higher within an approximate $\mu \pm 1.75\sigma$ window whereas $\mathcal{N}^{\Delta}$ will be higher outside this range due to its unbound nature. However, its implication on both **L** and **Ł** will depend on where this distribution lies relative to a given table, whether it is upon or below it. The difference can be significant, up to an approximate 2% difference in predicted probability in some cases Table 5-2. Overall, it is argued to err towards $\mathcal{N}$ unless compelled to do otherwise.

5.2.4. Inter-gas Correlations and Diagnosis

Motivation

Both estimation methods have proven viable for determining the probability of a given **L** or **Ł** while accounting for *Uncertainty* under the discussed assumptions. This Sub-Section demonstrates its natural extension in scope to include *Diagnosis* after the *Screening*. In addition, the significance of assuming inter-gas independency will be highlighted via a simple sensitivity analysis.

The motivation for incorporating the *Diagnosis* is that DGA-driven *Diagnostic Techniques* are based on the gas levels. Therefore, the *Uncertainty* can be expected to similarly affect the outputs of the *Diagnosis*. Although, work has been done elsewhere in literature looking at this topic such as [66], [120], the focus here is specifically propagating the *Uncertainty* through from the IEEE [1] *Screening* methodology which recommends the *Diagnosis* step be dependent on the *Screening* output. For this topic, only the *Duval Triangle 1* is considered, which was explained in Section 2.3. Other approaches could be incorporated following the same demonstrated approach. Fig. 5-7 explained the relevant methodology for this; if a given trial results in **Ł2–3**, it is then analysed via the *Duval Triangle 1* for a *Diagnostic Output*.

The motivation for incorporating inter-gas (in)dependency is that depending on the specific definitions used for *Measurement Uncertainty*, one can easily envisage

scenarios where multiple gases are affected by circumstances. For example, elevated TX temperatures, or a gas leak of the sampling syringe during the extraction for testing could be expected to affect more than one gas. The real challenge would be in quantifying and selecting appropriate parameters to explain such dependencies which may vary over time and between different gases. One very crude indicator is that the values in [1, Fig. A.9] show $P(Ł = 1) \approx 0.57$. If it is assumed the gases were independent, then it would approximate $P(L_g = 1) \sim 0.92$. Given that this is ostensibly the same dataset used to generate the limits, it may be expected that $P(L_g = 1) \lesssim 0.90$ based on $\mathbb{T}\mathbb{1}$ alone being set at the 90th percentile and **L1** requiring all $\mathbb{T}\mathbb{1}$–$\mathbb{4}$ to pass. This would therefore indicate there is a degree of inter-gas correlation. Otherwise, it may be expected for far fewer **Ł1**. Although too many assumptions would be needed to estimate the specific correlation value from this source to be reliable.

The methodology here is in line with the advice in [66, p. 25] to use the *Accuracy* metric. However, it could be argued the expected variability between gases measured by the same laboratory using the same method at the same time by presumably the same *Operator* will be less. For example, Fig. 3-8 indicates a tendency for a laboratory to either over-estimate or under-estimate the concentration levels for all gases. Therefore, perhaps a single metric such as *Accuracy* could be used for their collective "positioning" on the *Duval Triangle 1*, followed by a metric such as *Repeatability* as described in [68] for the positioning relative to one another. This is in contrast to using three instances of *Accuracy*. This is beyond the scope of this thesis which instead is simply demonstrating the potential impact of this parameter.

Input Data: M2

To output **Ł** and conditionally perform the *Diagnosis*, values for all the gases are required. Therefore, the **M2** dataset will be used which contains seven gases. **M2** uses contrived values, though the limits are representative of those included within the [1]'s tables. Table 5-3 contains the relevant sample values and limits.

For **M2**, only $\mathcal{N}$ will be considered, using both the N and $\mathbb{N}$ processing techniques previously discussed. Additionally, only one set of limits for $\mathbb{T}\mathbb{1}$–$\mathbb{4}$ will be considered, and instead, a new parameter representing inter-gas correlation will be explored. This parameter, ρ, will be set to either 0 or 1, with the former representing independence,

and the latter a linear dependence. This is very simplistic as it assumes a uniform relationship between all gases at all times and is intended solely to highlight the significance of the parameter. Although **L** and **Ł** will be provided for the **S1–3** *Protocols*, the *Diagnosis* will only be conducted on **S3**. For comparison, two different implementations will be used, $\boldsymbol{R}(\boldsymbol{\mathcal{N}})$ and $\boldsymbol{mR}(\boldsymbol{\mathcal{N}})$. These will be explained shortly. **L** would be unaffected by the methods used. All methods were tested to confirm this, and every comparable output across all methods were within ±0.05%. Therefore, only one instance of **L** results is included in the results shown later in Table 5-4.

*Table 5-3: Case Study Data **M2** with Limits*

Index	***Date**	$\bar{\mathbf{Y}}_i$ [PPM]							**Protocol**
i	[Days]	H_2	CH_4	C_2H_6	C_2H_4	$C2H_2$	CO	CO_2	*Sn*
1	720	80.00	50.00	90.00	50.00	^2.00	900	9000	1+2+3
2	576	40.00	51.00	65.00	40.00	^0.50	800	8000	2+3
3	432	67.00	54.00	94.00	51.00	^1.20	810	7710	3
4	288	45.00	54.00	82.00	42.00	^0.80	750	7320	3
5	144	42.00	43.00	78.00	37.00	^0.70	745	7350	3
6	0	40.00	40.00	70.00	32.00	^0.50	750	7500	3
Limits		H_2	CH_4	C_2H_6	C_2H_4	$C2H_2$	CO	CO_2	**Protocol**
τ_1 [PPM]		80.0	90.0	90.0	50.0	1.0	900	9000	1+2+3
τ_2 [PPM]		200.0	150.0	175.0	100.0	2.0	1100	12500	1+2+3
τ_3 [ΔPPM]		40.0	30.0	25.0	20.0	0.5	20	250	2+3
τ_4 [ΔPPM/Yr]		20.0	10.0	9.0	7.0	0.5	100	1000	3

***:** Calculated as days since oldest sample was taken.
^: Increased Uncertainty due to nearing LoD as per Equation (56).

Processing Methodology

It has been demonstrated that deriving the mathematical equations for these problems can be challenging, even if assuming independence. Therefore, no attempt is made to extend the derivations to include additional complications such as the *Diagnosis* or inter-gas dependencies. This also applies to the numerical integration method which would require said derived equations. Therefore, only the MCM is used as it has already been demonstrated to be suitably accurate and practical to implement with reasonable processing times. The general methodology is as previously shown in Fig. 5-7.

Inter-gas Correlations

Since there is no ground truth, two methods will be used for validation purposes, $\boldsymbol{mR}(\boldsymbol{\mathcal{N}})$ and $\boldsymbol{R}(\boldsymbol{\mathcal{N}})$. The primary method is the former, and it uses the "mvrnorm" function from the R package "MASS" [121] to draw one set of values for all gases for a given sample simultaneously. The relevant input argument is termed the "Sigma", Σ, and is described as a "positive-definite symmetric matrix specifying the covariance

matrix of all the variables" [121]. In this context, the previously mentioned parameter, ρ, was used to determine the appropriate Σ. This is done by rescaling a correlation matrix by the marginal variances as shown in Equation (124).

$$\Sigma = \begin{bmatrix} \sigma_{g=1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{g=G} \end{bmatrix} \begin{bmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sigma_{g=1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{g=G} \end{bmatrix}, \tag{124}$$

where the outer two are diagonal matrices of the gases' marginal standard deviations.

For the two special cases of $\rho = 0$ and $\rho = 1$, two simple alternative methods be used to validate the results, these are labelled the $\boldsymbol{R}(\mathcal{N})$ methods. For inter-gas independence, each can be sampled from appropriate $\mathcal{N}$ independently using the discussed "rnorm" function. For full correlation, a single gas can be sampled from an appropriate $\mathcal{N}$, and then each sample's quantile can be used to derive the equivalent value for each of the other gases. Equation (125) shows the equivalent operation.

$$y_i^g = \bar{Y}_i^g + \frac{\sigma_i^g \left(y_i^{g=1} - \bar{Y}_i^{g=1} \right)}{\sigma_i^{g=1}}. \tag{125}$$

where the superscripts represent a given gas, g, and **not** intended as exponents. Please note that Equations (62)–(64) deriving the probabilities **Ł1–3** assume inter-gas independency for validity and would not be otherwise applicable.

Duval Triangle 1 Diagnosis

Implementing this extension is trivial: if for a given trial, **Ł2–3** is output, then the relevant gas values will used to determine a *Diagnosis*. The sum of different *Diagnoses* can be compared to the number of trials to estimate the probabilities of each.

Results of Dataset M2

The results for **M2** dataset are shown in Table 5-4. Parameter of ρ does not affect **L1–3** and so only one set of these results are included for reference. Table 5-4 includes the outputs for **S1–3**, with the *Diagnosis* outputs based solely on **S3**. There are eight permutations of **Ł1–3**, these consider: the two implementation techniques, $\boldsymbol{mR}(\mathcal{N})$ and $\boldsymbol{R}(\mathcal{N})$; inter-gas independence and dependence, $\rho = 0$ and $\rho = 1$; and the previously discussed processing techniques of N and $\mathbb{N}$, respectively.

Processing Times

Table 5-4 shows there was no practical difference in outputs between the two processing techniques, N and $\mathbb{N}$. However, the former took approximately ~5 minutes whereas the latter took ~1 minute of processing time. This demonstrates its value even if only considering 6 samples. The difference would scale along with the total number of samples using in $\mathbb{T}4$. The implementation techniques also caused no practical difference in outputs. If assuming specifically $\rho = 0$ or $\rho = 1$, the $\boldsymbol{R}(\boldsymbol{\mathcal{N}})$ techniques were approximately 20% faster and so may be the preferred approach. However, for any other value or more complicated scenario, $\boldsymbol{mR}(\boldsymbol{\mathcal{N}})$ is needed.

Correlation Factor

Ł as evaluated via use of only measured values will not necessarily align with the estimated highest probability output. Indicatively, assuming a constant probability across each of the gases and inter-gas independence, $P(L_g = 1) \sim 0.91$ would result in an approximate $P(Ł = 1) \sim 0.5$. Conversely, $P(L_g = 3) \sim 0.09$ would result in an approximate $P(Ł = 3) \sim 0.5$. This means that a scenario having **L3** as the least likely outcome for all gases may still result in **Ł3** being the most likely outcome. If simplistically assuming ρ directly applied to **L**, then as $\rho \to 1$, $P(Ł = 1) := \min[P(L_g = 1)]$ and $P(Ł = 3) := \max[P(L_g = 3)]$. However, ρ is applied to the gas values and thus the results in Table 5-4 demonstrate less predictable results. Nevertheless, **Ł** for $\rho = 1$ is still closest to the worst-case **L**. It is therefore unsurprising that Table 5-4 highlights a significant difference between $\rho = 0$ and $\rho = 1$. Additionally, Table 5-4 shows how the conservative approach used in deriving **Ł** can greatly amplify the probability of **Ł3**, especially if considering inter-gas independence. While the magnitude of the impact of ρ varies by case, increased inter-gas correlation can be concluded to correspond to more optimistic outputs.

A conflict between the expectation of **Ł** depending on whether solely the measured value is used, or its estimated *Uncertainty* is included, can seem counter-intuitive given the *Measurement Uncertainty* was assumed unbiased centred about the measured values. The nonparametric operation of taking the worst-case **L** for **Ł** causes this. In a practical context, this can be considered problematic if the engineer is uncomfortable overriding the **Ł** output based on a crude estimate of *Measurement Uncertainty*. Depending on the relative importance given to this issue, one may opt for a less $\rho = 1$

to minimise the likelihood of it occurring, if considered critical, directly using the worst-case **L** may be a candidate metric to ensure no conflict.

*Table 5-4: Case Study Results of Data **M2** with Diagnoses*

	Probabilities, $P(X)$, [%]												
Protocol	***S1***			***S2***			***S3***			**^Duval Triangle Diagnosis**			
Per Gas*	$L1$	$L2$	$L3$	$L1$	$L2$	$L3$	$L1$	$L2$	$L3$	$X\|Ł2$		$X\|Ł3$	
H_2	50.0	50.0	0.0	42.6	57.4	0.0	42.6	52.0	5.4	–	–	–	–
CH_4	100.0	0.0	0.0	100.0	0.0	0.0	98.5	0.0	1.5	–	–	–	–
C_2H_6	50.0	50.0	0.0	40.0	60.0	0.0	38.8	45.7	15.5	–	–	–	–
C_2H_4	50.0	50.0	0.0	50.0	50.0	0.0	40.1	15.2	44.7	–	–	–	–
$C2H_2$	0.1	50.0	50.0	0.0	50.0	50.0	0.0	38.1	61.9	–	–	–	–
CO	50.0	49.8	0.2	17.8	82.0	0.2	16.4	61.8	21.9	–	–	–	–
CO_2	50.0	50.0	0.0	19.1	80.9	0.0	17.5	60.2	22.3	–	–	–	–
All Gas	**Ł1**	**Ł2**	**Ł3**	**Ł1**	**Ł2**	**Ł3**	**Ł1**	**Ł2**	**Ł3**	$T2\|Ł2$	$T3\|Ł2$	$T2\|Ł3$	$T3\|Ł3$
$\boldsymbol{\rho = 0}$													
$mR(\mathcal{N})_N$	0.0	49.9	50.1	0.0	49.9	50.1	0.0	10.1	89.9	75.7	24.3	56.5	43.5
$mR(\mathcal{N})_{\mathbb{N}}$	0.0	49.9	50.1	0.0	49.9	50.1	0.0	10.1	89.9	75.8	24.2	56.5	43.5
$R(\mathcal{N})_N$	0.0	49.9	50.1	0.0	49.9	50.1	0.0	10.1	89.9	75.8	24.2	56.5	43.5
$R(\mathcal{N})_{\mathbb{N}}$	0.0	49.9	50.1	0.0	49.9	50.1	0.0	10.1	89.9	75.8	24.2	56.5	43.5
$\boldsymbol{\rho = 1}$													
$mR(\mathcal{N})_N$	0.1	49.9	50.0	0.0	49.9	50.0	0.0	36.4	63.6	84.5	15.5	41.5	58.5
$mR(\mathcal{N})_{\mathbb{N}}$	0.1	50.0	50.0	0.0	50.0	50.0	0.0	36.5	63.5	84.5	15.5	41.5	58.5
$R(\mathcal{N})_N$	0.1	50.0	50.0	0.0	50.0	50.0	0.0	36.4	63.6	84.5	15.6	41.6	58.4
$R(\mathcal{N})_{\mathbb{N}}$	0.1	49.9	50.0	0.0	50.0	50.0	0.0	36.4	63.5	84.5	15.5	41.5	58.5

***:** All per-gas *DGA Status* outputs across all methods were within ±0.05%. $MV_N^{\rho=0}$ is shown.
^: No other output categories in example. Only applicable for combined *DGA Status*.
Note: Confidence Interval < ±0.03% ($n = 10^7$).

Diagnosis

The final aspect is the *Diagnosis* via the *Duval Triangle 1*. Care should be taken in the interpretation of the results in Table 5-4. The output is the probability of a given *Diagnosis* given *Measurement Uncertainty*. This is not the same as the probability of a given *Diagnosis* being true. Proximity to a border in the *Duval Triangle* may itself be considered a metric for confidence in its output, regardless of the *Measurement Uncertainty* being superimposed onto the ratio. This is because these borders are themselves somewhat arbitrary albeit data driven. Nevertheless, there can be argued a value in having an estimated probability of a given *Diagnosis* being 'repeatable' in an abstract sense. Its practical implication may be whether to prompt another sample, or whether there is already sufficient confidence in the output.

There is a significant difference in the probabilities of outputting **T2** or **T3** between $\rho = 0$ and $\rho = 1$. For the case of **Ł3**, the change in ρ from 0 to 1 caused a shift in the most likely output from **T2** to **T3**. In contrast, for **Ł2**, the change in ρ increased the

probability of **T3**. This demonstrates how the impact of ρ can be non-linear and difficult to predict. It is not clear what the practical implication would be in having differing *Diagnoses* depending on **Ł** other than to perhaps use them as inputs to a separate consolidation process. The IEEE methodology implies that *Diagnosis* is much more integral to **Ł3** than to **Ł2** and this could motivate the segregation. Otherwise, it seems an unnecessary complication in the outputs and instead perhaps having the combined probabilities of a given *Diagnosis* given **Ł2–3** would be better.

5.2.5. Demonstrative Example

This Sub-Section is intended to demonstrate a potential practical implementation. Therefore, the previously introduced **TX-D** will be used here, specifically, the "daily" variant. The raw gas values are shown in Fig. 4-16. The IEEE methodology [1] will be largely unchanged except for the following aspects. First, the limits used are those shown in Table 5-3. Though most are based on the IEEE tables, the limits for C_2H_2 are higher than those in [1] as was discussed in Section 5.1. Second, only the **S3** *Protocol* is considered here, and outputs prior to having sufficient samples for $\mathbb{T}4$ are simply ignored. Lastly, the modification discussed in Section 5.1 enforcing a 4-month minimum duration for $\mathbb{T}4$ will be applied.

Results of Dataset TX-D: Default Parameters

Initially, only $\rho = 0$ and $\rho = 1$ are explored. For the *Diagnosis*, only the combined probability given **Ł2–3** is included. Since the methodology has been separately validated, it is argued a practical use-case may use fewer trials for faster processing. Therefore, the trial count is lowered from 10^7 to $5x10^4$, this is approximately equivalent to a confidence interval of ±0.45% where $P = 0.5$ as per Equations (122) and (123). This allows for a processing time of the 708 samples in ~2 minutes.

Fig. 5-9 shows the outputs with **Ł** shown on the left column, and the *Diagnosis* outputs on the right. In this example, only the **T2** and **T3** *Diagnoses* were candidate potential outputs. The central band of outputs labelled "Expected Value" show the most likely outcome given the three considered scenarios. Scenario **A** assumes $\rho = 0$ and scenario **C** assumes $\rho = 1$. Scenario **B** assumes no *Measurement Uncertainty* such that the measured values are used directly. For scenarios **A** and **B** that have associated probabilities, the relative probabilities of a given outcome is shown above and below,

respectively. If considering the central band of outputs, Fig. 5-9 shows very similar behaviour across the three scenarios, with a slightly greater tendency for scenario **A** to conflict with **B** than would **C**, as should be expected. If comparing scenarios, **A** and **C**, their estimates for the probability of a given outcome seem quite symmetric. When considering the *Diagnosis* outputs, it is challenging to visually discern any meaningful difference. Fig. 5-9 shows that taken in context, the results are broadly very similar.

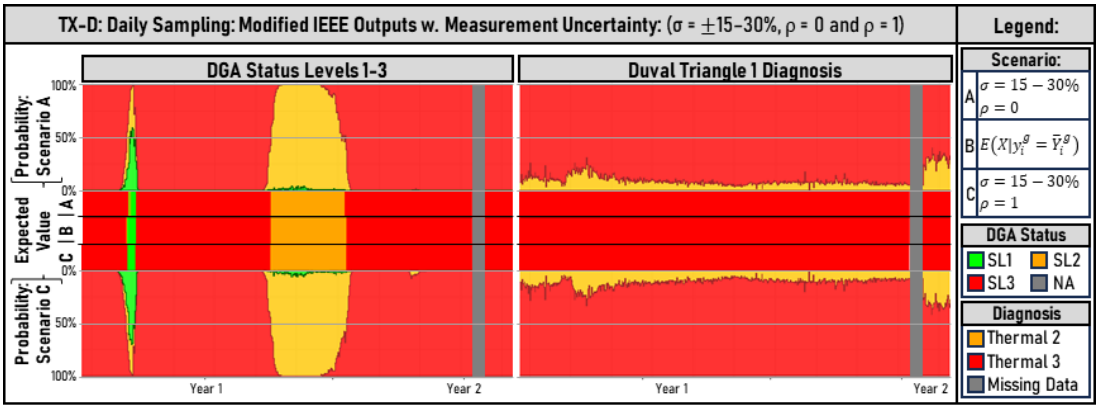


Fig. 5-9: Case Study Results of TX-D with Diagnoses: Scenarios A, B, and C

Processing Methodology for Derived Parameters:

The final aspect explored on this topic is regarding data-driven parameter selection. The methodology used is solely to explore the potential impact of using a more representative set of parameters for a given dataset as compared to the default values of the *IEC Specifications*. The derived estimates should not be considered "correct": but only that they may be closer to the "true" parameters than the default values.

Two parameters are estimated: the standard deviation and the correlation matrix describing the *Measurement Uncertainty*. It is being assumed that a measure of *Measurement Uncertainty* can be estimated as superimposed noise upon the gas levels given sufficient samples, i.e., the *Measurement Precision*. For simplicity, it is also being assumed that the *Measurement Uncertainty* is described by an *Absolute Uncertainty* rather than the *IEC Specification*'s *Relative Uncertainty*.

Thus far, the standard deviation of the *Measurement Uncertainty* was based on the *IEC Specification*, ranging between ±15% to ±30%, depending on the proximity to the LoD as per Equation (56). In lieu of this, a very simplistic method is implemented intended only to demonstrate the concept. *Measurement Uncertainty* is approximated based on the residuals from a 7-sample rolling median filter applied to **TX-D**. The distributions

are assumed unbiased. This is similar to the method suggested in [33, Sec. 10] to estimate a measure of an OLDGA's *Reproducibility*.

The tails of the empirical distributions are adjusted to mitigate the impact of perceived *Outliers*. It is argued that the ISO 5725 series generally identifies and removes *Outliers*, and the GUM series cautions there are far fewer relevant observations when estimating the tails of distributions empirically. The adjustment applied is that twice the average empirical distance from the median to either 2.5th percentile or the 97.5th percentile is used as an estimate of the 95% *Confidence Interval* of an estimated normal distribution. Then, the average empirical distance is used as the numerator in Equation (104), and 1.96 as the denominator, to estimate the final standard deviation.

Table 5-5 includes the empirical quantile values and standard deviation, and also the adjusted range and standard deviation for the gases. The impact of the adjustment depends on how closely the truncated portion of the empirical distribution resembles a normal distribution. Fig. 5-10 demonstrates this visually, with H_2 and C_2H_6 plotted on the left and right, respectively. In black is the empirical distribution, with its fitted normal distribution shown in red. The blue shows the adjusted fitted distribution based on truncating the *Outliers* as defined by values outside the vertically demarcated thresholds. H_2 had several outliers that if not removed caused a very wide fitted normal distribution whereas C_2H_6 was inherently closer to a normal distribution at the tails and so the adjustment had almost no effect.

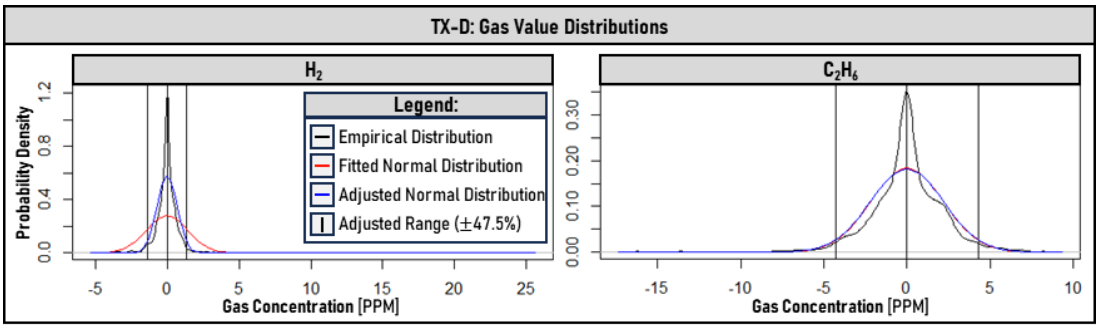


Fig. 5-10: Case Study TX-D Example Parameter Estimation Before and After Adjustment

Table 5-5: Case Study TX-D Parameter Estimation Before and After Adjustment

	Empirical Residuals				**Adjusted Residuals**	
	Quantile [PPM]			σ_g	**Range** [PPM]	σ_g
Gas	**2.5th**	**50th**	**97.5th**	[PPM]	**47.5th**	[PPM]
H_2	-1.5	0.0	1.2	0.69	1.35	1.44
CH_4	-2.4	0.0	1.9	1.30	2.15	1.10
C_2H_6	-4.3	0.0	4.3	2.17	4.30	2.19
C_2H_4	-1.8	0.0	1.6	0.82	1.70	0.87
$C2H_2$	-0.2	0.0	0.3	0.12	0.25	0.13
CO	-2.7	0.0	2.3	1.35	2.52	1.28
CO_2	-15.0	0.0	14.0	7.64	14.50	7.40

Table 5-6: TX-D Inter-Gas Correlation and Statistical Significance Estimates Before and After Adjustment

	Empirical Residuals [Correlation*Significance*^]							**Adjusted Residuals** [Correlation*Significance*^]						
Gas	**H_2**	**CH_4**	**C_2H_6**	**C_2H_4**	**C_2H_2**	**CO**	**CO_2**	**H_2**	**CH_4**	**C_2H_6**	**C_2H_4**	**C_2H_2**	**CO**	**CO_2**
H_2	1.00	*0.01*	*0.85*	*0.43*	*0.27*	*0.00*	*0.00*	1.00	*0.00*	*0.12*	*0.83*	*0.04*	*0.00*	*0.00*
CH_4	0.10	1.00	*0.00*	*0.07*	*0.59*	*0.00*	*0.00*	0.20	1.00	*0.00*	*0.00*	*0.46*	*0.00*	*0.00*
C_2H_6	-0.01	-0.18	1.00	*0.00*	*0.37*	*0.09*	*0.75*	-0.06	-0.31	1.00	*0.00*	*0.34*	*0.03*	*0.98*
C_2H_4	-0.03	0.07	0.48	1.00	*0.95*	*0.99*	*0.60*	0.01	0.11	0.32	1.00	*0.48*	*0.61*	*0.67*
$C2H_2$	-0.04	-0.02	0.03	0.00	1.00	*0.00*	*0.02*	0.08	0.03	0.04	0.03	1.00	*0.00*	*0.01*
CO	0.19	0.12	-0.06	-0.00	0.12	1.00	*0.00*	0.54	0.21	-0.09	0.02	0.11	1.00	*0.00*
CO_2	0.27	0.14	-0.01	0.02	0.09	0.61	1.00	0.45	0.16	-0.00	0.02	0.11	0.62	1.00

^: P-value of Pearson's product-moment correlation where null hypothesis is correlation is 0.

Previously, this Section explored constrained correlation matrix examples: keeping ρ constant for all gases and only varying between either a value of 0 or 1. However, the discussed "mvrnorm" function from the R package "MASS" [121] can accommodate intermediate values of ρ as well as different values of ρ for each inter-gas relationship. Therefore, using R's default "corr" function, correlation matrix was estimated. This was done using the residuals within the adjusted range, ignoring any sample that included a gas value that was truncated. For reference, the values obtained if using all samples are also included. As the output is a symmetric matrix about the diagonal, the top-right values are replaced with their respective p-values as per the Pearson's product-moment correlation test as implemented via R's default "corr.test" function, where a lower value indicates greater statistical significance [122]. For simplicity, all correlation values are taken naïvely as-is regardless of statistical significance. Table 5-6 tabulates the outputs.

Results of Dataset TX-D: Derived Parameters

Fig. 5-11 and Fig. 5-12 both follows the same format as Fig. 5-9. Fig. 5-11 continues using the *IEC Specification* and explores the impact of using the derived correlation matrix. Scenarios **A** and **B** are included for comparison, with scenario **D** along the bottom being the one using said correlation matrix shown on the right of Table 5-6.

Fig. 5-12 then uses the newly estimated standard deviation in its scenarios **E** and **F**. Scenario **E** on the top uses $\rho = 0$, and **F** on the bottom uses the estimated correlation matrix.

Fig. 5-11 is not implied as an ideal method and is only used to visualise the more subtle impact of the correlation matrix. Fig. 5-9 and Fig. 5-11 help confirm that the correlation matrix in the larger context is not highly impactful on the outputs as seen by the difficulty in differentiating the scenarios: **A**, **C**, and **D**. In contrast, the probabilities in Fig. 5-12 are noticeably different as compared to those in Fig. 5-9 and Fig. 5-11. The substantially tighter distributions driven by the smaller standard deviations result a more confident output. One potential added value that this implementation provides, as shown in the results, is greater granularity regarding trends. Prior to the expected **Ł** changing, the respective probabilities began changing first. This can provide forewarning of a given TX's **Ł** approaching a boundary. Though care should be taken to not conflate this with a forecasting tool predicting future *State of Health* (SoH) or *Probability of Failure* (PoF).

The *Duval Triangle 1* design appears robust, being less affected by *Measurement Uncertainty* in this case. This is somewhat surprising as inter-gas correlations should directly impact the relative ratios of said gases. However, this may be confounded by the correlations introduced by its metric design, and perhaps alternative forms, such as [123]'s simplex equivalent may be impacted further. Though this is not explored further in this thesis.

If sufficient data is available, estimating *Measurement Uncertainty* rather than relying on the default *IEC Specification* is recommended, as it can significantly impact outputs. For laboratory DGA, a laboratory's expected performance under [68]'s definitions of *Repeatability*, *Intra-Laboratory Reproducibility*, and *Accuracy* can be expected to be of great use. Without this, if taking samples annually, there may be insufficient data without pooling data from across multiple similar TXs. Estimating the inter-gas correlations is even more challenging as the link between it and the given performance metrics are more tenuous, e.g., attempting to infer based on differences between *Repeatability* and *Intra-Laboratory Reproducibility*. One consolation is that although the probability values were impacted, the most-likely outcome was rarely affected.

Nevertheless, as propagating *Measurement Uncertainty* is primarily to obtain the probabilities, available data should be utilised to adjust the parameters where possible.

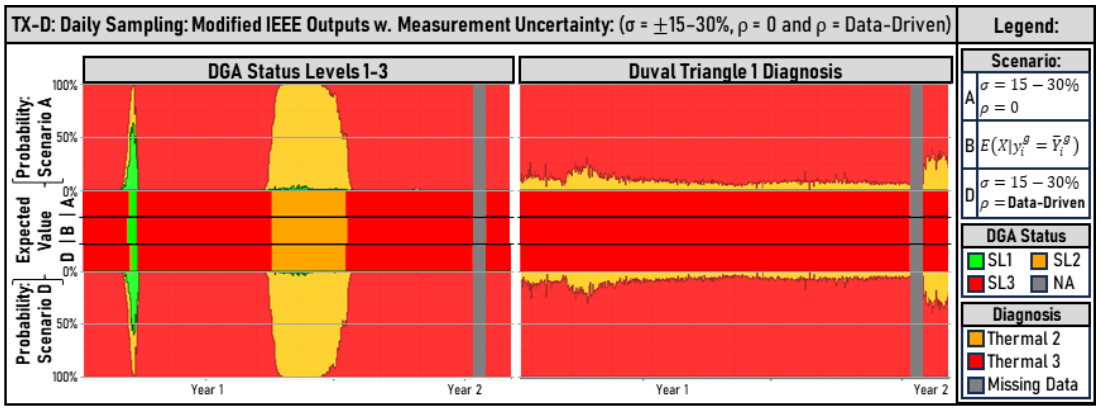


Fig. 5-11: Case Study Results of TX-D with Diagnoses: Scenarios A, B, and D

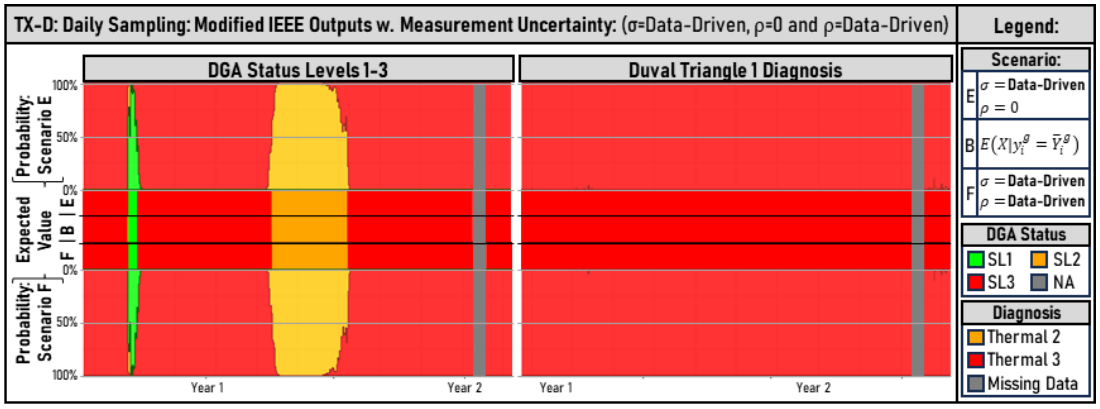


Fig. 5-12: Case Study Results of TX-D with Diagnoses: Scenarios E, B, and F

5.3. Conclusion

Section 5.1 concludes *Research Theme 1B* by contributing improvements to the IEEE C57.104-2019 methodology addressing potential barriers to practical deployment. Based on the literature reviewed in Chapter 3 and findings in Chapter 4, problematic edge cases of the methodology were identified and addressed through simple and intuitive improvements. A key issue identified with the methodology is its tendency to excessively flag samples at its maximal state, with the limits identified as a major contributor. Sub-Section 5.1.1 recommends the removal of all zero-based limits from the methodology to improve noise tolerance in practical deployment.

A key scenario identified in this thesis is the case where sampling rates are increased as a pre-emptive measure to closer inspect a TX. In this case, the act of increasing the sampling rate had an undesired "observer effect", affecting the outputs. Sub-Section

5.1.2 recommends a change to the derivation of the metric used in *Table 4*, $\mathbb{T}4$, of the methodology. The change is to remove the stipulation limiting the maximum number of samples used, and instead to introduce a stipulation limiting the minimum duration the samples must span. The recommended changes significantly improve the metric consistency in these cases.

Sub-Section 5.1.3 highlights the motivation for adjusting the derivations for the primary outputs of the methodology: the per-gas, **L**, and combined, **Ł**, *DGA Status*. The primary justification is the methodologies tendency to excessively output at its maximal level means that too many TXs are assigned the same rank—an effective *Screening* methodology should avoid this. Another minor change is rescaling it such that it follows the more conventional 0–1 TAI scale. This enhances compatibility with other indices, facilitating integration into practical deployments where multiple indices may be aggregated.

Sub-Section 5.1.4 recommends a linear interpolation between $\mathbb{T}1$ and $\mathbb{T}2$ to increase the granularity of the per-gas *DGA Status*, **L.** This simple change introduces a range of possible values to represent the equivalent of **L2**, allowing identification of TXs seemingly closer to escalating to the maximum **L3**. Sub-Section 5.1.5 increases the granularity of the combined *DGA Status*, **Ł**, by recommending the use of one of three candidate equations, Equation (59)–(61). These improvements aid prioritisation of TXs with otherwise same *DGA Status*, thus overcoming one of the primary issues identified with the IEEE C57.104-2019 methodology: excess **Ł3** flagging.

Section 5.2 concludes *Research Theme 2* by presenting a novel methodology to integrate *Measurement Uncertainty* into the IEEE C57.104-2019 methodology. Sub-Section 5.2.1 establishes the mathematical basis for the problem and Sub-Section 5.2.2 demonstrates the challenge in the algebraic application of the findings of Section 3.2 and Section 4.1 to the IEEE C57.104-2019 methodology. Sub-Section 5.2.3 then contributes two viable techniques for addressing this identified challenge. The first is via numerical integration and the second is a *Monte-Carlo Method* (MCM). These were cross validated to demonstrate their intended functionality. Furthermore, the implications to the changes to the derivation of $\mathbb{T}4$ of removing the maximum number

of samples were considered and addressed. The suggested technique to combine selected samples for $\mathbb{T}4$ was validated and shown to reduce computation times.

Sub-Section 5.2.4 demonstrates the extensibility of the proposed MCM methodology to incorporate inter-gas dependencies as well as the ability to cohesively propagate *Measurement Uncertainty* into the *Diagnostic* stage of the IEEE C57.104-2019 methodology. The viability and intended functionality of the proposed methodology was demonstrated via cross validation, using two other independent calculation techniques applicable to two specific scenarios: one of inter-gas independence, and one of fully linear inter-gas dependence. Sub-Section 5.2.4 also demonstrates that assuming inter-gas independence is conservative for the IEEE C57.104-2019 methodology and that full linear dependency across gases tends towards the outputs given by the worst-case gas. However, for the *Diagnosis*, the effect was shown to be difficult to predict.

Lastly, Sub-Section 5.2.5 concludes *Research Theme 2* by exploring the practical ramifications of the inclusion of *Measurement Uncertainty* into IEEE C57.104-2019. Real TX DGA data was used as a case study to consider two aspects. First, using the *IEC Specification*, outputs were compared between assuming inter-gas independence and assuming full linear dependence, where no *Measurement Uncertainty* acted as the control. The results indicated little difference in the outputs in the broader context. There was, as expected, a slightly greater tendency for the case assuming inter-gas independence to conflict with the outputs of the control of assuming no *Measurement Uncertainty*. Second, the impact of using *IEC Specifications* as compared to estimating alternative parameters based on the dataset was explored. It was again demonstrated that the value used for the inter-gas correlation had modest impacts on the outputs. However, the impact of assuming *IEC Specification*'s ±15–30% can be very significant, indicating priority should be allocated to refining this assumption prior deployment.

6. Conclusions and Further Work

Chapter Purpose and Structure

This Chapter summarises the primary conclusions drawn from the work presented in this thesis in Section 6.1, structured in accordance with the *Research Themes* introduced in Chapter 1. For further elaboration, refer to the relevant Chapters. Use of "IEEE methodology" and "IEC methodology" refer to IEEE C57.104-2019 and IEC 60599:2022, respectively. Section 6.2 concludes the thesis with a brief discussion on potential avenues for future work.

6.1. Conclusions

Chapter 2 presents a robust contextual background necessary to address the *Research Themes* via a comprehensive literature review of CIGRE Technical Brochures relevant to TX DGA for CMA. Each *Research Theme* is addressed individually.

Research Theme 1

Research Theme Scope

Research Theme 1A considers the impact the changes made to the IEEE methodology has to practical deployment. Sub-Section 3.1 reviews the IEEE methodology in detail, highlighting the key changes made in the new edition. The IEEE methodology is also compared to the IEC methodology. Furthermore, the NEI methodology outlined in [1 Annex F] was reviewed and compared to both its original publications ([3], [4]) and to LSA [5]; an industrially relevant methodology of potentially overlapping scope. Section 4.2 details the automated implementations of the *Screening* developed for each of the reviewed methodologies. These are used to assess the implications the changes to the IEEE methodology had on practical deployment via a detailed comparative analysis of outputs using case studies of real DGA data from 4 TXs.

Research Theme 1B uses the findings from *Research Theme 1A* to identify issues related to the practical deployment of the IEEE methodology. It then considers how, and to what extent, these issues may be addressed. Section 3.1 and Section 4.2 findings identify issues related to practical deployment of the IEEE methodology. Section 4.1 presents some simple conceptual experiments related to *Uncertainty* exploring potential practical significance. Section 4.1 findings help justify the proposed

improvements detailed in Section 5.1. Section 5.1 presents several improvements to the IEEE C57.104-2019 methodology attempting to address the identified issues.

Research Theme 1A Conclusions

Chapter 3 concludes that the IEEE C57.104-2019 and IEC 60599:2022 methodologies should not be considered to completely overlap in scope. It argues that the IEEE methodology is conceptually more focussed on solely providing a *Screening* output rather than it being indicative of *Fault Severity* as IEC methodology attempts. This narrower scope is more in line with literature in CIGRE regarding appropriate TAI design. Furthermore, Chapter 3 identifies that the IEEE methodology's primary outputs, the *DGA Status* (**Ł**) 1–3, do not align with the *Typical*, *Alert*, and *Alarm* scale of the IEC methodology. **Ł1** aligns best with *Typical*, **Ł2–3** lie closer to *Alert*, and the IEEE's *Extreme DGA Results* aligns best with *Alarm*. Another finding from Chapter 4 is that the appropriate selection of samples for IEEE's Table 4 (𝕋4) as well as the use of linear regression to calculate the gassing rate complicates the methodology's implementation as compared to the IEC methodology. This is exacerbated by the recommended use of a *Verification Sample*, which can potentially retrospectively alter outputs. This thesis presents a detailed interpretation to aid with developing an automated implementation of the IEEE methodology, expanded from the related publication: [75].

The application of the automated implementation of the IEEE methodology, developed for Chapter 4, identified an ambiguity in the intended use of the O_2/N_2 ratio. It is currently used to determine applicable limits; however, active gassing events can sufficiently change this ratio to alter the applicable limits, and thus the *DGA Status*. The practical significance of this finding is demonstrated using real TX DGA data in the case studies. This thesis recommends the guidance in IEEE C57.104-2019 be altered to explicitly state that the ratio should only be calculated during periods where no gassing is suspected to avoid this overlooked edge-case, and thus, improve the methodology's reliability. Another finding was that the guidance for using Duval Triangles 4 and 5 in IEEE C57.104-2019 is underspecified. It is currently undefined how to proceed when their outputs conflict, although this thesis does not propose a resolution.

Chapter 4 also demonstrates that the new metric derivation for the average gassing rate for $\mathbb{T}4$ in the IEEE methodology still tends to increase in sensitivity as the sampling rate is increased, although it successfully lessens the impact as compared to gassing rates calculated as per IEC 60599:2022. In practice, the introduced limitations on sample counts and timespans in $\mathbb{T}4$ can affect performance during periods of inconsistent sampling by either changing the applicable limit, or the applicability of $\mathbb{T}4$ in its entirety. In contrast, the new metric derivation for $\mathbb{T}3$ in the IEEE methodology was demonstrated to be resilient to most sampling rate changes as intended. However, it instead tends to reduce in sensitivity as sampling rates significantly increase.

Chapter 4 concludes that the IEEE methodology tends towards excessive TX flagging at its maximal output level, *DGA Status 3* (**Ł3**), when using default limits. Practical deployments expect a *Screening* methodology to effectively reduce the amount of candidate TXs to investigate and thus this presents a key barrier to adoption. A related recommendation of this thesis to benefit practical deployment is to introduce greater granularity to the output metric such that comparisons between TXs within the same level can be easier made.

The application of the automated implementation of the NEI method suggested in Annex F of IEEE C57.104-2019, developed for Chapter 4, demonstrated that it provides good contextual value and appears a direct improvement to more traditional aggregated metrics such as the T(D)CG. However, it also identifies insufficient guidance provided in [1, Annex F] for a practical implementation—a clear barrier to its adoption. The guidance provided in [1, Annex F] for the NEI is markedly less comprehensive than the original publications: [3] and [4]. However, these sources cannot easily be consolidated as their default values vary too much and no justification is provided in [1, Annex F] elaborating on why certain aspects were removed. The developed implementation demonstrates how a natural interpretation of the guidance results in unsatisfactory performance, identifying a need for future research for alternative interpretations. In particular, the aspects termed ‘auxiliary flags’ in Sub-Section 4.2.4 are identified as being underspecified.

The application of the automated implementation of the LSA [5] method, developed for Chapter 4, demonstrated that, as a *Screening* methodology, it is underspecified with

inadequate guidance particularly as the LSA [5] metric does not seem replicable from its source publication. Nevertheless, the metric is demonstrated, using real TX DGA data in the case studies, to perform very well in its ability to identify relevant changes in TX DGA at appropriate times. However, one identified weakness is that the LSA can be unpredictable if gas compositions change dramatically, such as after TX degassing, due to its emphasis on scaling the output based on CH_4 levels. It is concluded that the LSA metric is fundamentally too different to the NEI metric to be considered a functional equivalent. Rather, they function synergistically, with the latter capable of providing some of the context that the LSA may inherently lack.

These conclusions and case study results informs would-be users to the IEEE methodology of its expected behaviour, contextualised with comparisons to existing established methodologies. Furthermore, these conclusions provide the basis for addressing the practical deployment issues explored in *Research Theme 1B*.

Research Theme 1B Conclusions

Given the identified, and demonstrated, tendency of the IEEE methodology to excessively flag TXs, two avenues for improvement are recommended in Chapter 5. The first is to rectify causes of the excessive flagging, and the second is to increase the output granularity such that flagged TXs can be more readily ranked to facilitate their prioritisation for review. Together, these reduce the likelihood, and mitigate the consequence, of excessive flagging; effectively overcoming key barriers to practical deployment. In addition to these, Chapter 5 recommends a more natural 0–1 scaling *Transformer Assessment Index* (TAI) in place of the current 1–3 scale used for the both the per-gas (**L**) and combined (**Ł**) *DGA Status*, as this improves how readily it can be integrated with other TAIs. The means to do this are presented in Chapter 5.

On the topic of reducing flagging, Chapter 4 identifies the use of zero-values in default limits as a key driver to excessive flagging. Chapter 5 proposes a simple adjustment to raise these limits slightly to greatly improve practical performance. For example, to use *Limit of Detection* (LoD) or equivalent. Furthermore, as discussed, despite the changes to the calculation of the gassing rate for Table 4 (𝕋4) to reduce sensitivity to the sampling rate, there is still a strong undesired 'observer effect' introduced to the outputs when sampling rates are increased. This scenario arises when a TX is

considered suspect—the most critical period of assessment. Chapter 5 demonstrates that the stipulation capping the maximum number of samples for use in 𝕋4 to 6 is the primary cause. The basis of this argument being at a daily sampling rate, the 𝕋4's metric calculated on a 6-day span should not be considered comparable to the limits in its table. Either additional stratifications should be introduced to 𝕋4, or, as Chapter 5 recommends, a minimum duration should be stipulated to address this. This is demonstrated to improve metric consistency by reducing the distortion caused when increasing the sampling rate. These changes significantly contribute towards reducing flagging and directly improves the consistency in TX evaluation.

On the topic of increasing the *Screening* output granularity, the literature review concludes the decisive behaviour of the IEEE methodology's outputs is intended, and that the worst-case gas should remain heavily weighted. However, using solely the worst-case gas to represent the combined *DGA Status* (**Ł**) is identified in Chapter 5 as a key contributor to the excess **Ł3** outputs. Equation (61) from Sub-Section 5.1.5 is proposed as an alternative derivation to improve output granularity, provide unique outputs for each potential combination of per-gas *DGA Status* (**L**), and maintain the severe penalties based on the worst-case gas. Furthermore, the addition of a simple linear interpolation between 𝕋1 and 𝕋2 limits is proposed to effectively granulate **L2**, and potentially **Ł2**. These improvements, expanded from the related publication: [75], are demonstrated to increase the *Screening* output granularity.

Together, these changes proposed in Chapter 5 fulfil *Research Theme 1B* by enhancing the IEEE methodology's noise tolerance, metric consistency, and output granularity, effectively addressing highlighted deployment barriers whilst maintaining its perceived original intent.

Research Theme 2

Research Theme 2 Scope

Research Theme 2 examines the impact of changes made to the IEEE methodology concerning *Uncertainty*. This is concretised via the exploration of quantifying *Measurement Uncertainty* in IEEE methodology using a Standards-based approach. Section 3.2 reviews in detail the normative references of the IEEE methodology discussing the topic of *Uncertainty*, and in particular, *Measurement Uncertainty*. The

challenges associated with their practical application to the IEEE methodology was also discussed. Section 4.1 investigates the practical relevance of various factors potentially influencing *Uncertainty* via multiple simple conceptual experiments. Section 5.2 concludes *Research Theme 2* via in-depth analysis and application of a novel methodology to incorporate *Measurement Uncertainty* into the IEEE methodology.

Research Theme 2 Conclusions

Chapter 3 identifies a trend in the reviewed Standards-based literature towards the use of ISO/IEC Guide 98-3 (GUM), even if only via the ISO 5725 series 'top-down' approach. However, some historic values for *Uncertainty*, such as the *IEC Specification*'s ±15–30% *Accuracy*, have ambiguity regarding their correct interpretation in current contexts. Additional guidance is needed for the interpretation of the *IEC Specification* as well as for values for intermediate measures such *Intra-Laboratory Reproducibility*.

Uncertainty is mentioned as an important consideration in the IEEE methodology, but it lacks sufficient guidance to incorporate it. Chapter 3 identifies that applying guidance from GUM on ISO 5725 for application in IEC 60567's definition of *Measurement Uncertainty* as implicitly recommended via normative reference chain in IEEE C57.104-2019 is problematic. Chapter 3 concludes it is unclear how to overcome instances where the assumptions regarding the values for *Accuracy* and *Reproducibility* cannot be met concurrently.

The mathematical problem of quantifying *Measurement Uncertainty* in the IEEE methodology is presented in Sub-Section 5.2.1, expanded from the related publication: [47]. Even if assuming the gases within a sample and across samples are independent, it is still challenging to algebraically solve. Triangular distributions become overly cumbersome, requiring too many integration-by-parts, and $\mathcal{N}$ distributions become complex due to the non-linear relationship between the metrics used for $\mathbb{T}3$–4. Furthermore, it is argued that triangular distributions lose their primary appeal of intuitiveness due to the transformations necessary in calculating the outputs of the IEEE methodology.

Numerical estimation is recommended in Chapter 5 to address this identified gap, with detailed guidance on two practical methods: numerical integration and MCM. Furthermore, it is demonstrated that the inclusion of additional samples for $\mathbb{T}4$, as

recommended as part of *Research Theme 1B*, does not impede practical deployment if using the recommended mathematical simplification introduced in Chapter 5.

Chapter 5 also demonstrates that the recommended MCM method can readily incorporate inter-gas dependency, as well as extend the propagated *Measurement Uncertainty* into the *Diagnostic* stage of the IEEE methodology. It is demonstrated that assuming inter-gas independence is conservative in the IEEE methodology, with full linear dependency tending toward the same output as the worst-case gas. However, the effects of inter-gas dependencies on *Diagnostics* are difficult to predict and nonlinear.

Chapter 5 demonstrates that incorporating *Measurement Uncertainty* impacts the IEEE methodology's expected output significantly, irrespective of inter-gas dependencies. This is a consequence of using only the worst-case gas for its combined *DGA Status*. This may seem unintuitive to engineers and its implications should be considered prior deployment. Lastly, Chapter 5's preliminary results suggest minimal practical impact overall from the choice of inter-gas dependency, whether using the *IEC Specification* or other estimates for *Uncertainty*. In contrast, the impact of assuming *IEC Specification*'s ±15–30% for *Accuracy* is shown to be very significant. Refining this assumption prior to deployment should therefore be considered a priority over other factors.

Thus, *Research Theme 2* provides a detailed analysis on quantifying *Measurement Uncertainty* within the IEEE C57.104-2019 methodology. This includes identifying existing gaps in practical guidance within the normative references of IEEE C57.104-2019 and contributing, and demonstrating, a novel methodology to quantify the impact of *Measurement Uncertainty* via the propagation of probability distributions.

6.2. Further Work

Research Theme 1

This thesis neglected information generated by the Working Groups behind the IEEE and IEC methodologies, instead, heavily referencing literature generated by CIGRE's Working Groups to mitigate this shortcoming. Future work should look to incorporate them as it is their rationale and findings most relevant in shaping the methodologies they recommend. A challenge to be overcome, however, is that these documents are generally not publicly accessible. Furthermore, there is scope for a more comprehensive

rendition of a literature review on Standards related to *Uncertainty* in TX DGA CMA. In particular, GUM and ISO 5725 were too voluminous to review in detail and instead supplementary guidance documents of said Standards were relied upon. Future work should review the material more rigorously.

The automated implementations developed for the reviewed methodologies were simplified. For example, the recommended use of the *Verification Sample* in the IEEE methodology was neglected. Additionally, several comments in the documents related to specific situations were not implemented. These were detailed further in Sections 3.1 and 4.2. Establishing a more comprehensive implementation of the guidance outlined in IEEE methodology would add robustness to the recommendations. Furthermore, the inclusion of the previous IEEE methodology in future work may add helpful context. This was not done in this thesis due to its superseded status.

Many checks, such as the 'auxiliary flags' in Sub-Section 4.2.4, were difficult to integrate into an automated implementation. The implementation presented is too visually cluttered and should be improved upon in further work.

It remains challenging to automatically consider gas changes over a period of time to better isolate trends for *Diagnostics*. This thesis's chosen implementation of repeating *Diagnostics* twice, once with absolute values, and once with the delta from the greater of one month prior or the previous sample, is unsatisfactory. Further work could explore more sophisticated techniques to identify the appropriate time periods to isolate. Another improvement would be to incorporate a degree of interactivity into the implementation to allow the desired reference periods to be adjusted.

In Sub-Section 5.1.5 the topic of adding granularity to the *DGA Status* in the IEEE methodology was discussed. However, the probabilities estimated for a given *DGA Status* when incorporating *Measurement Uncertainty* presents an additional alternative approach that should be explored further.

The thesis provides an in-depth analysis of four TX case studies. Future work should extend the analysis to larger-scale studies, exploring a diverse range of conditions and macro-scale performance of the implementations. This will also enable the corroboration of the generality of the findings in this thesis.

Research Theme 2

It was shown that what qualifies as *Uncertainty* with the estimation of the average gassing rate is a nuanced topic that should be explored further. The implications of using OLDGA were particularly underexplored. It is easy to underestimate the complexity associated with interpreting what is represented by the slope coefficient of an estimated linear regression. Further research should include a more rigorous treatment of the topic exploring the relevance of violating assumptions typically required for validly using a linear regression, such as homoscedasticity of the residuals.

Section 5.2 utilises only the *Accuracy* metric which is potentially statistically flawed. When considering either the delta between two samples, or a gassing rate based on multiple samples, any shared influencing factors should not be duplicated. This means that for $\mathbb{T}3$–4, there ought to be a metric with a narrower range than that given by the *Accuracy* metric. For example, the *Intra-Laboratory Reproducibility* if assuming the samples are measured by the same laboratory. Alternatively, the *Accuracy* metric could instead be 'scaled' by a coefficient appropriate to the situation. This is especially the case for OLDGA, where short-term *Precision* can be expected to give much tighter intervals for $\mathbb{T}3$, and to a lesser extent, $\mathbb{T}4$. However, asserting both *Accuracy* and *Reproducibility* measures concurrently, as needed for $\mathbb{T}3$–4 in the IEEE methodology, can result in statistically incoherent assumptions that cannot be met. An example being if one gas sample was recorded as 0 ppm. Therefore, further research is required to establish the practical application to the IEEE methodology.

Section 5.2 overlooked the use of a uniform distribution, which should have been considered a viable option, given that the *IEC Specification* is a *Type B* source of *Uncertainty* with no specified *Coverage Factor*. Further research could consider including this distribution shape to the analysis.

Lastly, the work presented on inter-gas correlations represent only a preliminary introduction to a topic that has scope for much further elaboration. For example, inter-sample correlations could also be simultaneously considered. Additionally, more robust methodologies to estimate suitable parameters for the *Measurement Uncertainty* could be established.

7. References

[1] IEEE, 'Std C57.104-2019: IEEE Guide for the Interpretation of Gases Generated in Mineral Oil-Immersed Transformers', 2019.

[2] BSI, 'BS EN IEC 60599:2022: Mineral oil-filled electrical equipment in service – Guidance on the interpretation of dissolved and free gases analysis', 2022.

[3] F. Jakob, P. Noble, and J. J. Dukarm, 'A thermodynamic approach to evaluation of the severity of transformer faults', IEEE Transactions on Power Delivery, vol. 27, no. 2, pp. 554–559, Apr. 2012, doi: 10.1109/TPWRD.2011.2175950.

[4] F. Jakob and J. J. Dukarm, 'Thermodynamic Estimation of Transformer Fault Severity', IEEE Transactions on Power Delivery, vol. 30, no. 4, pp. 1941–1948, Aug. 2015, doi: 10.1109/TPWRD.2015.2415767.

[5] J. Lapworth, 'A Novel Approach (Scoring System) For Integrating Dissolved Gas Analysis Results Into A Life Management System', 2002 IEEE Internation Symposium on Electrical Insulation, USA, pp. 137–144, 2002.

[6] Western Power Distribution, 'Western Power Distribution's Distribution System Operability Framework', Jun. 2018. Available: westernpower.co.uk/DSOF.

[7] CIGRE WG A2.49, 'Condition assessment of power transformers', 2019.

[8] CIGRE WG A2.34, 'Guide for transformer maintenance', 2011.

[9] US Department of the Interior Bureau of Reclamation, 'Transformers: Basics, Maintenance, and Diagnostics'. 2005.

[10] EPRI, 'Power Transformer Maintenance and Application Guide', Sep. 2002. Available: epri.com/research/products/1002913.

[11] CIGRE WG A2.44, 'Guide on transformer intelligent condition monitoring (TICM) systems', 2015.

[12] BSI, 'BS EN IEC 60812:2018: Failure modes and effects analysis (FMEA and FMECA)'. 2018.

[13] BSI, 'BS ISO 18095:2018: Condition monitoring and diagnostics of power transformers', 2018.

[14] CIGRE WG A2.18, 'Life management techniques for power transformer', 2003.

[15] Ofgem, 'SPT.SHET_Network Asset Risk Annex (NARA)', 2018.

[16] CIGRE WG B3.12, 'Obtaining Value from On-Line Substation Condition Monitoring', 2011.

[17] Ofgem, 'Network Output Measures Methodology', 2017.

[18] G. Toman and R. Gazdzinski, 'Aging Management Guideline for Commercial Nuclear Power Plants - Power and Distribution Transformers', 1994.

[19] CIGRE WG A2.37, 'Transformer reliability survey', 2015.

[20] S. Tenbohlen et al., 'Results of a Standardized Survey about the Reliability of Power Transformers', Buenos Aires: CIGRE WG A2.37, Sep. 2017.

[21] IEEE, 'Std C57.143-2012: IEEE Guide for Application for Monitoring Equipment to Liquid-Immersed Transformers and Components', 2012. doi: 10.1109/IEEESTD.2012.6387561.

[22] CIGRE WG A2.27, 'Recommendations for condition monitoring and condition assessment facilities for transformers', 2008.

[23] BSI, 'BS ISO 55000:2014: Asset management – Overview, principles and terminology', 2014.

[24] D. Wright, A. Kelly, and A. Stuart, 'Electricity Transmission Network Output Measures Methodology', 2016.

[25] L. Esserman, Y. Shieh, and I. Thompson, 'Rethinking Screening for Breast Cancer and Prostate Cancer', JAMA, vol. 302, no. 15, pp. 1685–1692, 2009.

[26] K. P. Murphy, Probabilistic Machine Learning: Advanced Topics. MIT Press, 2023. Accessed: Oct. 02, 2022. Available: probml.ai.

[27] CIGRE WG A2.55, 'Life extension of oil filled transformers and shunt reactors', 2022.

[28] BSI, 'BS EN 60812:2006: Analysis techniques for system reliability. Procedure for failure mode and effects analysis (FMEA)', 2006.

[29] BSI, 'BS EN ISO/IEC 17000:2020: Conformity assessment - Vocabulary and general principles', 2020.

[30] ISO/IEC, 'ISO/IEC Guide 98-3:2008, Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995)', 2008.

[31] J. S. Bendat and G. P. Allan, 'Random Data: Analysis and Measurement Procedures', 4th ed. Wiley Series in Probability and Statistics, 2010.

[32] ISO/IEC, 'ISO/IEC Guide 98-6:2021, Uncertainty of measurement — Part 6: Developing and using measurement models', 2021.

[33] CIGRE D1.01 (TF15), 'Report on gas monitors for Oil-Filled electrical equipment', 2010.

[34] S. Bell, 'Good Practice Guide No. 11 The Beginner's Guide to Uncertainty of Measurement', Teddington, Mar. 2001.

[35] BSI, 'BS ISO 5725-2:2019: Accuracy (trueness and precision) of measurement methods and results. Basic method for the determination of repeatability and reproducibility of a standard measurement method', 2019.

[36] BSI, 'BS ISO 5725-3:2023: Accuracy (trueness and precision) of measurement methods and results. Intermediate precision and alternative designs for collaborative studies', 2023.

[37] BSI, 'BS ISO 5725-4:2020: Accuracy (trueness and precision) of measurement methods and results. Basic methods for the determination of the trueness of a standard measurement method', 2020.

[38] BSI, 'BS ISO 5725-5:1998: Accuracy (trueness and precision) of measurement methods and results. Alternative methods for the determination of the precision of a standard measurement method', 1998.

[39] BSI, 'BS ISO 5725-6:1994: Accuracy (trueness and precision) of measurement methods and results. Use in practice of accuracy values', 1994.

[40] BSI, 'BS ISO 5725-1:2023: Accuracy (trueness and precision) of measurement methods and results. General principles and definitions', 2023.

[41] N. Roderick, J. A. Little, and D. B. Rubin, 'Statistical Analysis with missing Data', 3rd ed. Wiley, 2019.

[42] S. van Buuren, 'Flexible Imputation of Missing Data', 2nd ed. Chapman & Hall, 2018.

[43] ISO/IEC, 'ISO/IEC Guide 98-1:2009, Uncertainty of measurement — Part 1: Introduction to the expression of uncertainty in measurement', 2009.

[44] Wikipedia contributors, 'Riemann sum', Wikipedia, The Free Encyclopedia. Available: wikipedia.org/w/index.php?title=Riemann_sum.

[45] ISO/IEC, 'ISO/IEC Guide 98-3:2008/Suppl.1:2008, Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995) — Supplement 1: Propagation of distributions using a Monte Carlo method', 2011.

[46] Wikipedia contributors, 'Quasi-Monte Carlo method', Wikipedia, The Free Encyclopedia. Available: wikipedia.org/w/index.php?title=Quasi-Monte_Carlo_method.

[47] M. Hosseini and B. G. Stewart, 'Propagating Uncertainty using IEEE Std C57.104-2019 Dissolved Gas Analysis Methodology for Transformers', in 23rd International Symposium on High Voltage Engineering (ISH 2023), Glasgow: IET, 2023, pp. 698–704.

[48] JCGM WG 1, 'An introduction to the "Guide to the expression of uncertainty in measurement" and related documents', 2009.

[49] D. Rediansyah, R. A. Prasojo, Suwarno, and A. Abu-Siada, 'Artificial intelligence-based power transformer health index for handling data uncertainty', IEEE Access, vol. 9, pp. 150637–150648, 2021, doi: 10.1109/ACCESS.2021.3125379.

[50] N. Meinshausen, 'Quantile Regression Forests', Journal of Machine Learning Research, vol. 7, pp. 983–999, 2006.

[51] J. I. Aizpurua, V. M. Catterson, B. G. Stewart, S. D. J. McArthur, B. Lambert, and J. G. Cross, 'Uncertainty-Aware Fusion of Probabilistic Classifiers for Improved Transformer Diagnostics', IEEE Trans Syst Man Cybern Syst, vol. 51, no. 1, pp. 621–633, Jan. 2021, doi: 10.1109/TSMC.2018.2880930.

[52] W. Feilhauer and E. Handschin, 'Interpretation of Dissolved Gas Analysis Using Dempster-Shafer's Theory of Evidence', in 9th International Conference on Probabilistic Methods Applied to Power Systems, Stockholm: IEEE, Jun. 2006, pp. 1–6.

[53] R. E. Neapolitan, 'Learning Bayesian Networks'. New Jersey: Prentice-Hall, 2004.

[54] C. E. Lin, J. M. Ling, C. L. Huang, and S. M. Member, 'Expert System for Transformer Fault Diagnosis Using Dissolved Gas Analysis', 1993.

[55] L. A. Zadeh, 'Fuzzy Sets', Information and Control, vol. 8, pp. 338–353, 1965.

[56] P. Hajek, L. Godo, and F. Esteva, 'Fuzzy logic and probability', 2013, doi: 10.48550/arXiv.1302.4953.

[57] L. A. Zadeh, 'Discussion: Probability Theory and Fuzzy Logic Are Complementary Rather Than Competitive', vol. 37, no. 3, pp. 271–276, 1995.

[58] P. Baraldi, L. Podofillini, L. Mkrtchyan, E. Zio, and V. N. Dang, 'Comparing the treatment of uncertainty in Bayesian networks and fuzzy expert systems used for a human reliability analysis application', Reliab Eng Syst Saf, vol. 138, pp. 176–193, 2015, doi: 10.1016/j.ress.2015.01.016.

[59] Y.-C. Huang and H.-C. Sun, 'Dissolved Gas Analysis of Mineral Oil for Power Transformer Fault Diagnosis Using Fuzzy Logic', 2013.

[60] J. C. Drotos, J. W. Porter, and R. Stebbins, 'Dissolved Gas Analysis of Transformer Oil', 1996.

[61] Martin Heath Cote Associates, 'Maintenance of High Voltage Transformers', London, 1989.

[62] CIGRE WG D1/A2.47, 'Advances in DGA interpretation', 2019.

[63] J. Golarz, 'Understanding Dissolved Gas Analysis (DGA) Techniques and Interpretations', Electric Energy T&D Magazine, pp. 31–36, Nov. 2015.

[64] CIGRE WG D1.32, 'DGA in non-mineral oils and load tap changers and improved DGA diagnosis criteria', CIGRE, 2010.

[65] M. Hosseini and B. G. Stewart, 'ANRC 11-3: Advanced Transformer Health Monitoring: Project Report 01', Glasgow, Jan. 2020.

[66] M. Duval and J. J. Dukarm, 'Improving the reliability of transformer gas-in-oil diagnosis', IEEE Electrical Insulation Magazine, vol. 21, no. 4, pp. 21–27, Jul. 2005, doi: 10.1109/MEI.2005.1489986.

[67] BSI, 'BS EN IEC 60475:2022: Method of sampling insulating liquids', 2022.

[68] BSI, 'BS EN IEC 60567:2011: Oil-filled electrical equipment. Sampling of gases and analysis of free and dissolved gases', 2011.

[69] CIGRE WG D1/A2.47, 'DGA monitoring systems', 2019.

[70] CIGRE TF 15-01-07, 'New techniques for dissolved gas analysis', Electra, vol. 198, pp. 20–27, 2001.

[71] Michel Duval, 'New Techniques for Dissolved Gas-in-Oil Analysis', IEEE Electrical Insulation Magazine, pp. 6–15, Mar. 2003.

[72] M. Duval, 'Calculation of DGA Limit Values and Sampling Intervals in Transformers in Service', IEEE Electrical Insulation Magazine, vol. 24, no. 5, pp. 7–13, Sep. 2008.

[73] D. Lamontagne, 'Utilizing Piecemeal Linear Approximation and Harmonic Regression to Analyze Power transformer Insulating Oil O-Line Gas Samples', in TechCon North America, 2010.

[74] IEEE, 'PES Transformers Committee: Standards Subcommittee'. Accessed: Jun. 10, 2023. Available: transformerscommittee.org/subcommittees/standardssc/.

[75] M. Hosseini, B. G. Stewart, M. Kearns, and N. Torenvliet, 'Construction of a Transformer DGA Health Index Based on DGA Screening Processes', in Annual Report - Conference on Electrical Insulation and Dielectric Phenomena, CEIDP, IEEE, Oct. 2020, pp. 391–394. doi: 10.1109/CEIDP49254.2020.9437537.

[76] BSI, 'BS EN IEC 60599:2015: Mineral oil-filled electrical equipment in service – Guidance on the interpretation of dissolved and free gases analysis', 2015.

[77] CIGRE JTF D1.01/A.211, 'Recent developments in DGA interpretation', 2006.

[78] IEEE 'Std C57.104-2008: IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers', 2009.

[79] ISO/IEC, 'ISO/IEC Guide 98-1:2024, Guide to the expression of uncertainty in measurement — Part 1: Introduction', 2024.

[80] ISO/IEC, 'ISO/IEC Guide 98-4:2012, Uncertainty of measurement — Part 4: Role of measurement uncertainty in conformity assessment', 2012.

[81] ISO/IEC, 'ISO/IEC Guide 98-3:2008/Suppl.2:2011, Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995) — Supplement 2: Extension to any number of output quantities', 2008.

[82] ISO/IEC, 'ISO/IEC Guide 99:2007, International vocabulary of metrology — Basic and general concepts and associated terms (VIM)', 2007.

[83] BSI, 'BS ISO 10576:2022: Statistical methods - Guidelines for the evaluation of conformity with specified requirements', 2022.

[84] BSI, 'PD ISO/TR 13587:2012: Three statistical approaches for the assessment and interpretation of measurement uncertainty', 2012.

[85] BSI, 'PD IEC Guide 115:2023: Application of measurement uncertainty to conformity assessment activities in the electrotechnical sector', 2023.

[86] BSI, 'BS EN ISO/IEC 17025:2000: General requirements for the competence of testing and calibration laboratories', 2000.

[87] BSI, 'BS ISO 3534-1:2006: Statistics - Vocabulary and symbols: General statistical terms and terms used in probability', 2006.

[88] BSI, 'BS ISO 3534-2:2006: Statistics - Vocabulary and symbols: Applied statistics', 2006.

[89] BSI, 'BS ISO 21748:2017: Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty evaluation', 2017.

[90] BSI, 'DD ISO/TS 21749:2005: Measurement and uncertainty for metrological applications - Repeated measurements and nested experiments', 2005.

[91] BSI, 'BS EN 60567:2005: Oil-filled electrical equipment - Sampling of gases and of oil for analysis of free and dissolved gases - Guidance', 2005.

[92] BSI, 'BS EN 60567:1993: Guide for the sampling of gases and of oil from oil-filled electrical equipment and for the analysis of free and dissolved gases', 1993.

[93] D. A. Armbruster and T. Pry, 'Limit of Blank, Limit of Detection and Limit of Quantitation', Clin Biochemist Reviews, vol. 29, pp. S50–S52, 2008.

[94] BSI, 'BS 5497-1:1987: Precision of test methods - Part 1: Guide for the determination of repeatability and reproducibility for a standard test method by inter-laboratory tests', 1987.

[95] Wikipedia contributors, 'Mean absolute percentage error', Wikipedia, The Free Encyclopedia. Available: wikipedia.org/w/index.php?title=Mean_absolute_percentage_error.

[96] BSI, 'BS ISO 5725-1:1994: Accuracy (trueness and precision) of measurement methods and results. General principles and definitions', 1994.

[97] M. Hosseini and B. G. Stewart, 'ANRC 11-3: Advanced Transformer Health Monitoring: Project Report 02', Glasgow, May 2020.

[98] M. Hosseini and B. G. Stewart, 'ANRC 11-3: Advanced Transformer Health Monitoring: Project Report 02: Addendum', Glasgow, Jun. 2020.

[99] I. Farrance and R. Frenkel, 'Uncertainty of Measurement: A Review of the Rules for Calculating Uncertainty Components through Functional Relationships', 2012.

[100] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, 'A basic introduction to fixed-effect and random-effects models for meta-analysis', Res Synth Methods, vol. 1, no. 2, pp. 97–111, Apr. 2010, doi: 10.1002/jrsm.12.

[101] G. Solon, S. J. Haider, and J. M. Wooldridge, 'What Are We Weighting For?', Source: The Journal of Human Resources, vol. 50, no. 2, pp. 301–316, 2015.

[102] T. S. Breusch and A. R. Pagan, 'A Simple Test for Heteroscedasticity and Random Coefficient Variation', Econometrica, vol. 47, no. 5, pp. 1287–1294, Sep. 1979.

[103] A. Zeileis and T. Hothorn, 'Diagnostic Checking in Regression Relationships', R News, vol. 2, no. 3, pp. 7–10, 2002.

[104] Wikipedia contributors, 'Kronecker delta', Wikipedia, The Free Encyclopedia. Available: wikipedia.org/w/index.php?title=Kronecker_delta.

[105] H. Wickham, 'ggplot2: Elegant Graphics for Data Analysis'. Springer-Verlag New York, 2016. Available: ggplot2.tidyverse.org.

[106] M. R. Smith, 'Ternary: An R Package for Creating Ternary Plots', Comprehensive R Archive Network, 2017, doi: 10.5281/zenodo.1068996.

[107] Wikipedia contributors, 'Error function', Wikipedia, The Free Encyclopedia. Available: wikipedia.org/w/index.php?title=Error_function.

[108] Wikipedia contributors, 'Analytic function', Wikipedia, The Free Encyclopedia. Available: wikipedia.org/w/index.php?title=Analytic_function.

[109] W. Viechtbauer, 'Conducting meta-analyses in R with the metafor package', J Stat Softw, vol. 36, no. 3, pp. 1–48, 2010, doi: 10.18637/jss.v036.i03.

[110] National Bureau of Standards, 'Tables of the Bivariate Normal Distribution Function and Related Functions', vol. 50. Applied Mathematics, 1959.

[111] M. Lipow, N. Mantel, and J. W. Wilkinson, 'Query 2: The Sum of Values from a Normal and a Truncated Normal Distribution (Continued)', Technometrics, vol. 6, no. 4, pp. 469–471, 1964, doi: doi.org/10.2307/1266101.

[112] R. Carnell, 'Triangle: Distribution Functions and Parameter Estimates for the Triangle', Dec. 2022. Available: CRAN.R-project.org/package=triangle.

[113] H. W. Borchers, 'pracma: Practical Numerical Math Functions', 2022, CRAN: 2.4.2. Available: CRAN.R-project.org/package=pracma.

[114] P. J. Davis and P. Rabinowitz, Methods of Numerical Integration. New York: Dover Publications, 1984.

[115] L. F. Shampine, 'MATLAB Program for Quadrature in 2D', in Proceedings of Applied Mathematics and Computation, 2008, pp. 266–274.

[116] Wikipedia contributors, 'Gauss-Kronrod quadrature formula', Wikipedia, The Free Encyclopedia. Available: wikipedia.org/w/index.php?title=Gauss%E2%80%93Kronrod_quadrature_formula&oldid=1192134857.

[117] L. D. Brown, T. T. Cai, and A. DasGupta, 'Interval Estimation for a Binomial Proportion', Statistical Science, vol. 16, no. 2, pp. 101–117, May 2001.

[118] A. Agresti and B. A. Coull, 'Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions', Am Stat, vol. 52, no. 2, pp. 199–126, 1998.

[119] W. T. Scherer, T. A. Pomroy, and D. N. Fuller, 'The triangular density to approximate the normal density: decision rules-of-thumb', Reliab Eng Syst Saf, vol. 82, no. 3, pp. 331–341, 2003.

[120] B. G. Stewart and J. I. Aizpurua, 'Uncertainty Analysis of Two Gas Measurement DGA Ratios for Improved Diagnostics Applications', in 2022 IEEE International Conference on High Voltage Engineering and Applications, 2022. doi: 10.1109/ICHVE53725.2022.9961491.

[121] W. N. Venables and B. D. Ripley, 'Modern Applied Statistics with S', 4th ed. Springer, 2002.

[122] Wikipedia contributors, 'Pearson correlation coefficient', Wikipedia, The Free Encyclopedia. Available: wikipedia.org/w/index.php?title=Pearson_correlation_coefficient.

[123] J. Dukarm, Z. Draper, and T. Piotrowski, 'Diagnostic simplexes for dissolved-gas analysis', Energies (Basel), vol. 13, no. 23, Dec. 2020, doi: 10.3390/en13236459.

[124] Wikipedia contributors, 'PERT distribution', Wikipedia, The Free Encyclopedia. Available: wikipedia.org/w/index.php?title=PERT_distribution.

Annex A: IEEE Tables

Tables A-1 to A-4 are tables 1 to 4 from [1, Sec. 6]. These are here for convenience but should be considered in context of the entirety of the guidance given in [1].

Table A-1: 90th percentile gas concentrations as a function O_2/N_2 ratio and age in µL/L (ppm)

		O_2/N_2 Ratio ≤ 0.2				O_2/N_2 Ratio > 0.2			
		Transformer Age in Years				Transformer Age in Years			
		Unk.	1 – 9	10 – 30	>30	Unk.	1 – 9	10 – 30	>30
Gas	Hydrogen (H_2)	80	75		100	40	40		
	Methane (CH_4)	90	45	90	110	20	20		
	Ethane (C_2H_6)	90	30	90	150	15	15		
	Ethylene (C_2H_4)	50	20	50	90	50	25	60	
	Acetylene (C_2H_2)	1	1			2	2		
	Carbon monoxide (CO)	900	900			500	500		
	Carbon dioxide (CO_2)	9000	5000	10000		5000	3500	5500	

NOTE—During the data analysis, it was determined that voltage class, MVA, and volume of mineral oil in the unit did not contribute in significant way to the determination of values provided in Table 1.

Source: Table 1 from [1]

Table A-2: 95th percentile gas concentrations as a function O_2/N_2 and age in µL/L (ppm)

		O_2/N_2 Ratio ≤ 0.2				O_2/N_2 Ratio > 0.2			
		Transformer Age in Years				Transformer Age in Years			
		Unk.	1 – 9	10 – 30	>30	Unk.	1 – 9	10 – 30	>30
Gas	Hydrogen (H_2)	200	200			90	90		
	Methane (CH_4)	150	100	150	200	50	60		30
	Ethane (C_2H_6)	175	70	175	250	40	30	40	
	Ethylene (C_2H_4)	100	40	95	175	100	80	125	
	Acetylene (C_2H_2)	2	2		4	7	7		
	Carbon monoxide (CO)	1100	1100			600	600		
	Carbon dioxide (CO_2)	12500	7000	14000		7000	5000	8000	

NOTE—During the data analysis, it was determined that voltage class, MVA, and volume of mineral oil in the unit did not contribute in significant way to the determination of values provided in Table 2.

Source: Table 2 from [1]

Table A-3: 95th percentile values for absolute level change between successive laboratory DGA samples in µL/L (ppm)

		Maximum µL/L (ppm) variation between consecutive laboratory DGA samples	
		O_2/N_2 Ratio ≤ 0.2	**O_2/N_2 Ratio > 0.2**
Gas	Hydrogen (H_2)	40	25
	Methane (CH_4)	30	10
	Ethane (C_2H_6)	25	7
	Ethylene (C_2H_4)	20	
	Acetylene (C_2H_2)	Any Increase	
	Carbon monoxide (CO)	250	175
	Carbon dioxide (CO_2)	2500	1750

NOTE—Contribution of voltage class, MVA, and volume of mineral oil in the unit was not studied for Table 3 as they have not been retained for Table 1 and Table 2. Data was insufficient to study age influence. Source: Table 3 from [1]

Table A-4: 95th percentile values from multi-points (3-6 points) rate analysis of laboratory DGA samples with all gas levels below Table 1 values, in µL/L (ppm/year)

		Maximum µL/L/year (ppm/year) rate in function of the period between first and last point of the laboratory DGA series (3 to 6 samples)			
		O_2/N_2 Ratio ≤ 0.2		**O_2/N_2 Ratio > 0.2**	
		Period between first and last point of the series			
		4–9 Months	10–24 Months	4–9 Months	10–24 Months
Gas	Hydrogen (H_2)	50	20	25	10
	Methane (CH_4)	15	10	4	3
	Ethane (C_2H_6)	15	9	3	2
	Ethylene (C_2H_4)	10	7	7	5
	Acetylene (C_2H_2)	Any increasing rate		Any increasing rate	
	Carbon monoxide (CO)	200	100	100	80
	Carbon dioxide (CO_2)	1750	1000	1000	800

NOTE—Contribution of voltage class, MVA, and volume of mineral oil in the unit was not studied for Table 4 as they have not been retained for Table 1 and Table 2. Data was insufficient to study age influence. Source: Table 4 from [1]

Annex B: Limit Selection

Defining the appropriate limits for the integrals described in Sub-Section 5.2.1 is convoluted due to the large number of permutations. Rather than define each uniquely and explicitly, Equations (B.1)–(B.27) represent the logic for passing or failing a given table in relation to the limit selection of a given sample. The relevant combination would be used, taking either the minimum for the upper limit, or maximum for the lower limit. For example, $\check{\Psi}_{1,T1}$ in Equation (B.1) represents the lower integral limit for sample Y_1 to pass $\mathbb{T}\mathbb{1}$, i.e., the same as described for $\check{\gamma}_1$ for Equation (88). As another example, $\hat{\gamma}_{1,L2}$ from Equation (92) is the upper limit of the integral for Y_1 where $\mathbb{T}\mathbb{1}$ is passing and $\mathbb{T}\mathbb{3}$ is failing, causing **L2**. This would therefore be represented by the minimum of $\hat{\Psi}_{1,T1}$ and $\neg\hat{\Psi}_{1,T3}$, as shown in Equations (B.3) and (B.14), respectively.

Equations (B.1)–(B.4): Limits related to $\mathbb{T}\mathbb{1}$

$$\check{\Psi}_{1,T1} = \check{Y}_1, \quad (B.1) \qquad \hat{\Psi}_{1,T1} = min\{\hat{Y}_1|\tau_1\}, \quad (B.3)$$

$$\neg\check{\Psi}_{1,T1} = max\{\check{Y}_1|\tau_1\}, \quad (B.2) \qquad \neg\hat{\Psi}_{1,T1} = \hat{Y}_1. \quad (B.4)$$

Equations (B.5)–(B.8): Limits related to $\mathbb{T}\mathbb{2}$

$$\check{\Psi}_{1,T2} = \check{Y}_1, \quad (B.5) \qquad \hat{\Psi}_{1,T2} = min\{\hat{Y}_1|\tau_2\}, \quad (B.7)$$

$$\neg\check{\Psi}_{1,T2} = max\{\check{Y}_1|\tau_2\}, \quad (B.6) \qquad \neg\hat{\Psi}_{1,T2} = \hat{Y}_1, \quad (B.8)$$

Equations (B.9)–(B.15): Limits related to $\mathbb{T}\mathbb{3}$

$$\check{\Psi}_{1,T3} = \check{Y}_1, \quad (B.9) \qquad \hat{\Psi}_{1,T3} = min\{\hat{Y}_1|\hat{Y}_2 + \tau_3\}, \quad (B.13)$$

$$\neg\check{\Psi}_{1,T3} = max\{\check{Y}_1|\check{Y}_2 + \tau_3\}, \quad (B.10) \qquad \neg\hat{\Psi}_{1,T3} = \hat{Y}_1, \quad (B.14)$$

$$\check{\Psi}_{2,T3} = max\{\check{Y}_2|y_1 - \tau_3\}, \quad (B.11) \qquad \hat{\Psi}_{2,T3} = \hat{Y}_2, \quad (B.15)$$

$$\neg\check{\Psi}_{2,T3} = \check{Y}_2. \quad (B.12)$$

Equations (B.16)–(B.27): Limits related to $\mathbb{T}\mathbb{4}$

$$\check{\Psi}_{1,T4} = \check{Y}_1, \quad (B.16)$$

$$\hat{\Psi}_{1,T4} = min\left\{\begin{array}{l}\hat{Y}_1, \\ \dfrac{\tau_4 - \left(c_2 \times \boxed{\tilde{Y}_2}\right) - \left(c_{\mathbb{N}} \times \hat{Y}_{\mathbb{N}}\right)}{c_1}\end{array}\right\| \boxed{\tilde{Y}_2} = \begin{cases}\hat{Y}_2, & x_2 < \bar{x}_{N,} \\ 0, & x_2 = \bar{x}_{N,} \\ \check{Y}_2, & \bar{x}_N < x_{2,}\end{cases} \quad (B.17)$$

$$\neg\check{\Psi}_{1,T4} = max\left\{\begin{array}{l}\check{Y}_1, \\ \dfrac{\tau_4 - \left(c_2 \times f_2\left(\boxed{\tilde{Y}_2}\right)\right) - \left(c_{\mathbb{N}} \times f_{\mathbb{N}}\left(\hat{Y}_{\mathbb{N}}\right)\right)}{c_1}\end{array}\right\| \boxed{\tilde{Y}_2} = \begin{cases}\check{Y}_2, & x_2 < \bar{x}_{N,} \\ 0, & x_2 = \bar{x}_{N,} \\ \hat{Y}_2, & \bar{x}_N < x_{2,}\end{cases} \quad (B.18)$$

$$\neg\hat{\Psi}_{1,T4} = \hat{Y}_1, \tag{B.19}$$

$$\check{\Psi}_{2,T4} = min\left\{\begin{matrix}\check{Y}_2, \\ \boxed{\tilde{Y}_2}\end{matrix}\middle\|\boxed{\tilde{Y}_2} = \begin{cases}\dfrac{\tau_4 - (c_1 \times y_1) - \left(c_{\mathbb{N}} \times f_{\mathbb{N}}\left(\hat{Y}_{\mathbb{N}}\right)\right)}{c_2}, & x_2 < \bar{x}_N, \\ \check{Y}_2, & x_2 \geq \bar{x}_N, \end{cases}\right. \tag{B.20}$$

$$\hat{\Psi}_{2,T4} = min\left\{\begin{matrix}\hat{Y}_2, \\ \boxed{\tilde{Y}_2}\end{matrix}\middle\|\boxed{\tilde{Y}_2} = \begin{cases}\hat{Y}_2, & x_2 \leq \bar{x}_N, \\ \dfrac{\tau_4 - (c_1 \times y_1) - \left(c_{\mathbb{N}} \times f_{\mathbb{N}}\left(\hat{Y}_{\mathbb{N}}\right)\right)}{c_2}, & x_2 > \bar{x}_N, \end{cases}\right. \tag{B.21}$$

$$\neg\check{\Psi}_{2,T4} = max\left\{\begin{matrix}\hat{Y}_2, \\ \boxed{\tilde{Y}_2}\end{matrix}\middle\|\boxed{\tilde{Y}_2} = \begin{cases}\check{Y}_2, & x_2 \leq \bar{x}_N, \\ \dfrac{\tau_4 - (c_1 \times y_1) - \left(c_{\mathbb{N}} \times f_{\mathbb{N}}\left(\check{Y}_{\mathbb{N}}\right)\right)}{c_2}, & x_2 > \bar{x}_N, \end{cases}\right. \tag{B.22}$$

$$\neg\hat{\Psi}_{2,T4} = min\left\{\begin{matrix}\check{Y}_2, \\ \boxed{\tilde{Y}_2}\end{matrix}\middle\|\boxed{\tilde{Y}_2} = \begin{cases}\dfrac{\tau_4 - (c_1 \times y_1) - \left(c_{\mathbb{N}} \times f_{\mathbb{N}}\left(\check{Y}_{\mathbb{N}}\right)\right)}{c_2}, & x_2 < \bar{x}_N, \\ \hat{Y}_2, & x_2 \geq \bar{x}_N, \end{cases}\right. \tag{B.23}$$

$$\check{\Psi}_{\mathbb{N},T4} = \check{Y}_{\mathbb{N}}, \tag{B.24}$$

$$\hat{\Psi}_{\mathbb{N},T4} = min\left\{\begin{matrix}\hat{Y}_{\mathbb{N}}, \\ \dfrac{\tau_4 - (c_1 \times y_1) - (c_2 \times y_2)}{c_{\mathbb{N}}},\end{matrix}\right. \tag{B.25}$$

$$\neg\check{\Psi}_{\mathbb{N},T4} = max\left\{\begin{matrix}\check{Y}_{\mathbb{N}}, \\ \dfrac{\tau_4 - (c_1 \times y_1) - (c_2 \times y_2)}{c_{\mathbb{N}}},\end{matrix}\right. \tag{B.26}$$

$$\neg\hat{\Psi}_{\mathbb{N},T4} = \hat{Y}_{\mathbb{N}} \tag{B.27}$$

where $\boxed{\tilde{Y}_2}$ represents a variable that is defined within a given Equation. Depending on whether the sampling time, x_2, of Y_2 is less than, equal to, or greater than the average sampling time, its relative impact to β_1 changes. This means that an increase in Y_2 can either increase, decrease, or have no impact on β_1. Thus, the need for the variable $\boxed{\tilde{Y}_2}$. In contrast, $Y_{\mathbb{N}}$ will have to be equivalent to lying on the opposite side of Y_1.