CIGRE-408

2021 CIGRE Canada Conference & Expo
Toronto, Ontario, October 25-28 2021

Forecasting Near-Term Failure of Transformers Using Reliability Statistics on Dissolved Gas Analysis

Z.H. DRAPER[1], J.J. DUKARM
Delta-X Research, Inc.
Canada

SUMMARY

Dissolved gas analysis (DGA) is widely used for periodic screening of transformers. DGA promises an affordable and non-invasive way to identify energized transformers at risk of failure. This symptom-based strategy allows for targeted physical testing and prioritization for asset maintenance. Often, DGA results are scored and put into a 'health index' to do a so-called 'condition-based' maintenance assessment. However, dissolved gases in oil are only an indirect symptom of a problem and do not truly indicate condition. Therefore, the gases are not a condition-based measurement. This means timely and accurate interpretation of fault-related gassing is critical for a preventative asset management program. However, various degrees of industry experience with DGA have led the IEEE C57.104-2019 to go as far as to say DGA is "more of an art than a science". So how well do DGA fault severity assessment methods actually work?

One of the purposes of DGA is to predict a near-term transformer failure. Historically, IEEE and IEC DGA fault severity methods have used 90th (or higher) percentiles of gas concentration, increments, and/or rates of change based on large amounts of DGA data. The prevailing view has been that larger gas values represent a worse health status for the transformer. These conventional methods primarily rely on a few case studies to prove their validity rather than a robust population study of in-service transformer failures. Recently, a new method of dissolved gas analysis developed by the authors, called Reliability-based DGA (R-DGA), uses failure data and reliability statistics to derive failure rate curves from DGA data [1, 2]. The failure rate curves obtained in that way show that some of the highest rate of failure with gassing occurs before a 90th percentile level is reached. A previously non-gassing transformer that starts gassing might quickly be at higher risk of near-term failure than a transformer that has accumulated a large amount of gas over time.

To determine how well commonly used DGA severity assessment methods perform, we compiled a database of transformer DGA histories from five electric utilities with 15,239 operating transformers and 307 failure cases. We calculated the status codes for all of the transformers using conventional methods such as IEC 60599-2015, IEEE C57.104-2008, and IEEE C57.104-

[1]**zhdraper@deltaxresearch.com**

2019 [3–5]. When we compared those limits-based methods against R-DGA [1, 2] we found that DGA interpretation by means of reliability statistics performed better at predicting near-term transformer failure. We used screening-test statistics and optimization curves on the DGA classification results to quantify and compare the relative capability of the various DGA assessment methods to predict near-term failure. Such statistical methods provide an objective rubric for improving DGA interpretation methods and the preventative asset management programs that depend on them.

KEYWORDS

DGA, Power transformers, Risk assessment.

1 UTILITY FLEET DGA DATA

In order to compare the ability of DGA interpretive methods to predict near-term failure, a large database of DGA data containing failure cases in a realistic context was required. DGA data for entire fleets of transformers were collected from 5 North American electric utility companies. For each failure case, there was a date of failure and a description of the root cause of failure. Post-failure DGA samples were excluded because they would not be predictive of the past event. Failure cases for which the root cause was not predictable by DGA, such as vandalism or wildlife, was excluded. To be included, each case was required to include at least three DGA samples, with at least one DGA sample less than two years prior to failure. In most cases there was a sample within a year prior to failure, given the typical industry practice of a yearly sampling frequency. In total there were 15,239 transformers that were considered to be in service as of the last sample, and there were 307 transformers that failed in service. A best faith effort was made to include only examples of DGA-predictable failure, despite the variability of detail in the root cause of failure descriptions. We hope that the results presented here will motivate electric utilities to keep good records of failure cases to make tangible improvements in DGA interpretation.

2 DGA INTERPRETATION METHODS

2.1 Conventional Methods

Standards bodies such as IEC and IEEE provide guidelines for the interpretation of DGA data. Their general methodology is to compute percentile limits (such as the 90th and 95th), based on a large DGA database, to determine which gas values are very atypical to observe. Then a procedure is provided to apply the limits to assign a status level. We represent those status levels by numeric codes, usually 1-4 or 1-3, where the highest number indicates the greatest degree of concern. In this study, we tried 3 conventional methods as close to a 'cookbook' application as we could, however, the standards are ambiguous about some specific details.

For IEEE C57.104-2008 [4], we used the values from Table 1 as concentration limits for defining 4 status codes and ignored the footnote which says to use those limits only if one sample is available. This has been the general industry practice, and otherwise the method would not be practical. For IEEE C57.104-2019 [5] we used all of the concentration, rate of change, and increment limits from Tables 1-4 following the flow chart of logic in Figure 2 to establish 3 status codes. The application of rate of change calculations is ambiguous, but we calculated up to 4 potential observed rates of change, varying based on the number of possible samples within a given time interval, and tested if any rate of change exceeded a defined limit. For IEC 60599-2015 [3], the so-called 'alert' and 'alarm' levels are not explicitly defined and instead the document defines a range of 'typical' values at the 90th percentile. For the purposes of this study, we use the logic defined in IEC 60599-2015, but take limits from CIGRE TB 771 in Tables C.7 and C.8 [6]. Above-typical values are considered an 'alert' level, and the 'Pre-Failure' level [7] is considered an 'alarm' level. We will consider the IEC 'alarm' level to be status code 3 and 'alert' to be status code 2. Because of ambiguity in the IEEE and IEC guidelines, variations in the interpretation may be considered more or less reasonable and may impact the final results of this comparison. This paper's purpose is to convey that the comparative methodology described in Section 3 can help define a more rigorous and accurate DGA interpretation method by validation testing possible variations.

2.2 Reliability-based DGA

Instead of just assuming that atypically large gas values represent a worse condition or status for the transformer, it is possible to use statistical survival analysis with DGA and failure data to calculate the risk associated with gassing. A key simplification is to base the the statistical analysis on normalized energy intensity (NEI) indices [8, 9]. One, representing fault energy "cracking" the insulating oil, is based on the standard heats of formation from oil of the four hydrocarbon gases methane, ethane, ethylene, and acetylene. It is called NEI-HC. The other, representing fault energy charring paper insulation, is based on the standard heats of formation from cellulose of CO and CO_2. It is called NEI-CO.

As discussed in [1] and illustrated in Figure 2 of that paper, reliability statistics can be applied to derive a survival probability model and a corresponding hazard rate curve for NEI-HC and for NEI-CO. The hazard rate curve in particular represents the risk of short-term failure associated with increasing NEI-HC or NEI-CO. The main indicator of failure risk associated with a transformer's active gassing is called "hazard factor" (HF), with units "percent failures per year" [2]. It is calculated by multiplying the most recent rate of increase of each NEI by the most recent value of the corresponding hazard rate. Then those two quantities are added to obtain a combined HF. HF thus combines information about the rate of gassing (NEI units per year) with the NEI risk (percent failures per additional unit of NEI) to obtain HF as an incremental failure rate (percent failures per year). For quantitative assessment of the risk associated with DGA results, R-DGA ranks transformers according to the continuous quantity HF.

To represent R-DGA results in terms of discrete risk levels analogous to those provided by the limits-based IEEE and IEC methods, we can say that an extreme risk level (4) is for transformers with HF above its 90th percentile value over a very large population. The high risk level (3) is for transformers with HF > 0.05, but can also include a more moderate level of risk of HF < 0.05. The lower risk level (2) is for transformers with HF of zero (in other words no active gassing) but it does have a prior history of gassing.

3 COMPARING METHODS

3.1 General Remarks

Since in-service transformer failures are relatively rare, it is fair to say that the main purpose of DGA screening is to identify transformers that are behaving abnormally (by producing fault gases) and potentially might require some kind of investigation or maintenance. Of course it is especially important that, in particular, transformers in danger of failing should be included among those identified by DGA as requiring urgent attention. The DGA interpretation method comparison undertaken for the purposes of this study focuses on their ability to predict near-term transformer failure. A more general comparison of DGA methods could be carried out by defining a condition less radical than failure but still a well defined objective of DGA screening. For example, diagnosing cooling system inefficiencies and maintenance.

3.2 Screening Test Statistics

The results of analyzing the transformers based on the 4 different DGA interpretation methods can be seen in Figure 1. In order to have an ideal method for flagging transformers at risk of failure for an asset management program, fewer assets deemed to be operating should be flagged, while most of the assets that failed in service should be flagged as a high degree of

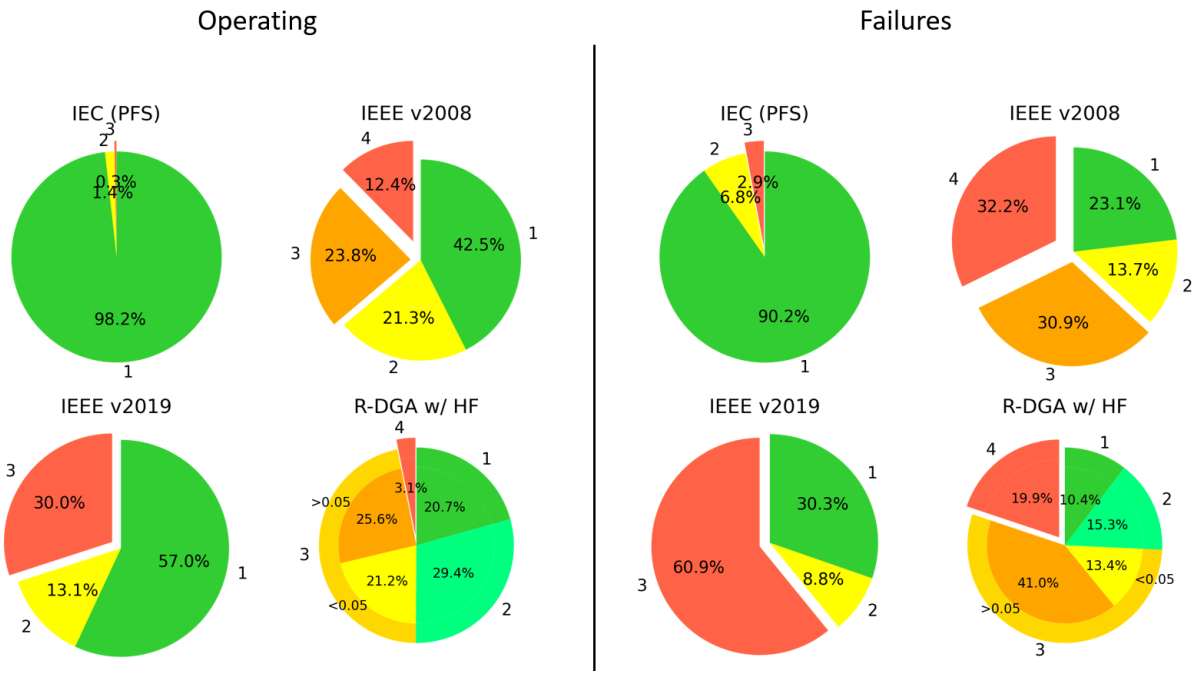

Figure 1: The relative fraction of status codes for the operating transformers (left) and transformers which failed in-service (right). Red-Orange-Yellow represent higher to lower risk categories under each DGA assessment method. Light green-Green represents transformers determined to not be at risk of failing in the near-term. Ideally, the number of transformers flagged by DGA in the left side should be low where as the right side with failure cases should be high. That would represent a method that is specifically flagging potential failures with a minimal number of false negatives and positives.

risk. Flagging assets unnecessarily represents wasted effort and physical testing to confirm a non-existent problem in the transformer. Failing to flag an asset before it fails on the other hand, represents lost revenue from an unplanned outage and extra costs associated with a catastrophic event.

In order to objectively characterize how well each method performs, we use screening test statistics. Definitions and equations for these statistics can be found in [10] and [11]. This kind of statistics are widely used in the medical community to determine how well a medical test performs at flagging a serious condition within a patient (e.g. a fecal immunochemical test (FIT) for colon cancer [12]). It involves splitting the group of subjects by a known condition and comparing against their test results. In this case, the condition being tested for is whether or not the transformer is about to fail and the test result is the result code of the DGA interpretation method. We also include a 'coin flip' method to compare against a random classifier to see if DGA has any predictive value. Also, we compute statistics for a realistic 'gold standard' which is a hypothetical test that has a 95% accuracy rate, but is unbiased towards the true condition.

From the results in Table 1 we learn quite a few things. First, all DGA interpretation methods are able to do better than a random coin flip, but are not ideal or efficient at specifically predicting failure as one might hope. The lower the status codes, the higher the true positive rate. Yet this comes at a cost of additional false positives. A methods efficiency at balancing this trade off can be seen using the diagnostic odds ratio. The 95% confidence intervals of DOR determine which methods are statistically equivalent in their efficiency given the available data. The top three methods of equivalent DOR is the IEC method and R-DGA status code 4. However, it can bee seen that R-DGA status 4 has a higher true positive rate, meaning it will flag more of the potentially costly transformer failures, with the same efficiency as IEC. Overall, the IEC

Test Method	Prev.	*TPR*	*FPR*	*PPV*	*DOR*	*DOR* 95% CI
Gold Standard	1.97%	95.0%	5.0%	27.68%	361	[215, 606]
IEC: 3	1.97%	2.9%	0.3%	15.52%	9.36	[4.56, 19.23]
R-DGA: 4	1.97%	17.9%	2.4%	13.16%	8.94	[6.56, 12.19]
IEC: 2	1.97%	9.8%	1.77%	10.0%	6.03	[4.06, 8.95]
R-DGA: 3, HF > 0.05	1.97%	60.9%	28.7%	4.10%	3.87	[3.07, 4.88]
IEEE 2019: 3	1.97%	60.9%	30.0%	3.93%	3.64	[2.89, 4.59]
IEEE 2008: 4	1.97%	32.3%	12.4%	4.99%	3.37	[2.64, 4.3]
IEEE 2019: 2	1.97%	69.7%	43.0%	3.16%	3.04	[2.38, 3.89]
IEEE 2008: 3	1.97%	63.2%	36.2%	3.40%	3.03	[2.4, 3.83]
R-DGA: 3, HF < 0.05	1.97%	74.3%	49.9%	2.91%	2.9	[2.24, 3.75]
IEEE 2008: 2	1.97%	76.9%	57.5%	2.62%	2.46	[1.88, 3.21]
Fair Coin Flip	1.97%	50.0%	50.0%	1.97%	1	[0.8, 1.25]

Table 1: Screening test statistics for multiple risk levels of four DGA interpretation methods and two hypothetical methods for comparison. The rows are sorted by the diagnostic odds ratio (DOR). The 95% confidence intervals on the DOR illustrate which methods are statistically equivalent in diagnostic efficiency. 'Prev.' is the prevalence of failure. TPR is the true positive rate or sensitivity. FPR is the False positive rate. PPV is the positive predictive value.

method does not flag very many of the failure cases, although when the IEC method is triggered, there is strong positive predictive value in IEC status codes. The positive predictive value for each status codes tells us how much we can believe that the transformer has a potential failure condition given that test is positive. In Bayesian statistics, this would represent the *a posteriori*. Theses PPV values could be used to update the predicted probability of failure for an asset management system given one of the DGA status codes comes up positive. The higher status codes are generally more predictive, but none achieve better than 50%, suggesting there is weak correlation of the condition with the test. This is consistent with the idea that DGA is only the starting point to prioritize further physical testing and determine the true condition of a transformer. IEEE 2019 status code 3 only has a positive predictive value of 3.93% which is only about twice that of the background prevalence of failure in the population. IEEE 2019 allows for an undefined 'Extreme' DGA category which will clearly be necessary to increase the predictive value of the method and find the worst cases.

3.3 Optimization Curves

In addition to screening test statistics, there are optimization curves that can determine how well a classifier is able to rank a condition overall. Originally developed for radar systems in WWII, an ROC curve is more commonly used in modern day machine learning algorithms to compare classification methods using the TPR and FPR [13]. The TPR and FPR can vary with a ranking metric like the DGA status codes. R-DGA provides a rank scoring metric with HF which means it is able to better classify the risk of each individual asset giving it a more refined ROC curve compared to the other methods. The ROC curve for DGA interpretation methods can be seen in Figure 2. Generally the best classifier is the one which has a larger area under the curve. This area represents the probability that a randomly chosen transformer will be correctly ranked higher than a randomly chosen transformer without the condition being tested for. In general, R-DGA has the largest areas under the curve (0.70), while IEEE 2008 and 2009 come in second with similar values (0.66), and IEC is last (0.54). The precision recall curve (See Figure 3) is another optimization curve which is similar to the ROC curve but is sometimes useful when there is a large imbalance in the number of true condition cases [14].

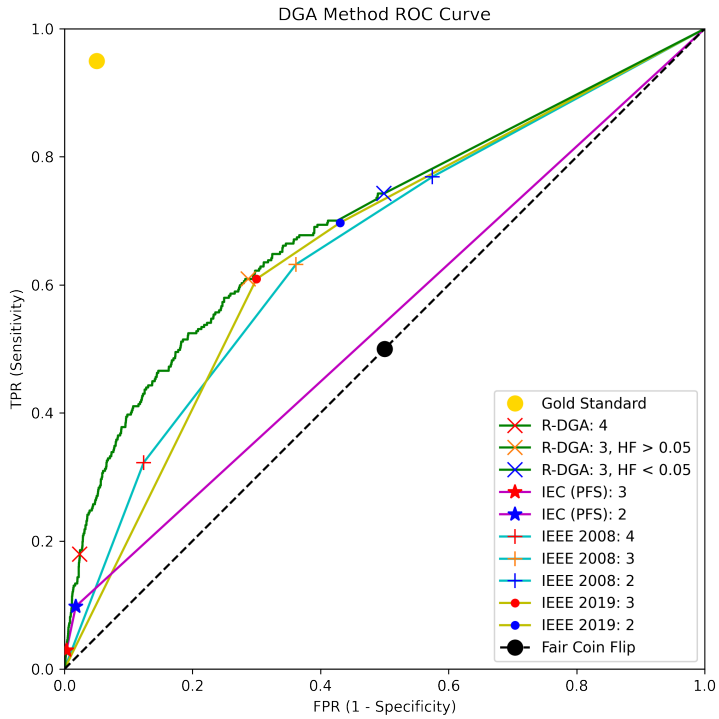

Figure 2: ROC curves for DGA assessment methods. The y-axis represents the true positive rate while the x-axis represents the false positive rate. The path along each curve from the lower left to upper right represent a high to low risk of failure implied by each method. The black dashed line represents a random coin flip classifier, which has no diagnostic value. Curves which reach closer to the upper left corner (near the yellow dot representing the hypothetical 'gold standard' test) generally represent better classification methods. The areas under each curve represent the overall effectiveness of the respective methods.

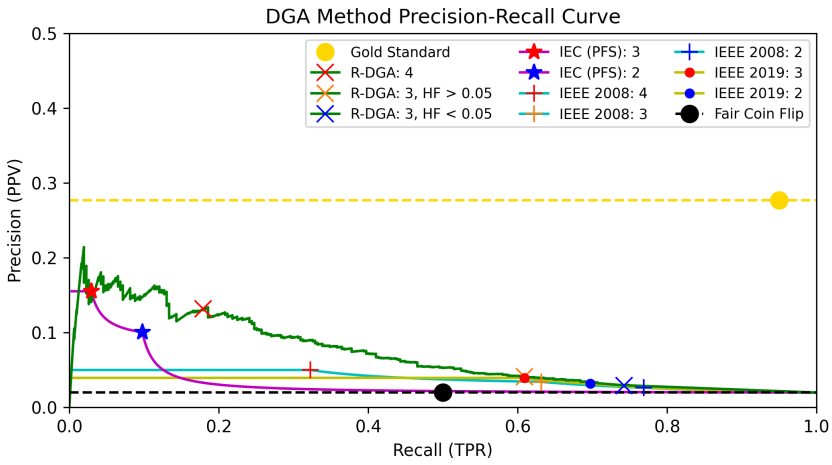

Figure 3: DGA method Precision-Recall (P-R) curve. When condition positive and negative are strongly imbalanced, a Precision-Recall curve can be more informative than an ROC curve [14]. The plot shows the precision (PPV) as a function of recall (TPR). The hypothetical 'coin flip' and 'gold standard' lines are plotted for comparison.

4 CONCLUSIONS

By using objective metrics such as screening test statistics and optimization curves, we can compare DGA interpretation methods. Based on this comparison methodology, we find that the methods described in [1, 2] for Reliability-based DGA can out perform the conventional methods developed by IEC or IEEE at flagging transformers prior to failure. Furthermore, the positive predictive values tabulated in this paper can be used to update a Bayesian model on the chances of failure given that a particular DGA method comes up positive. The comparative methodology in this paper is applied to DGA for the first time and illustrates a more rigorous way of constructing a DGA interpretation method. With the evident success on 5 fleets of transformers, more data should be collected to further refine DGA interpretation for the risk assessment of transformers.

BIBLIOGRAPHY

[1] J. J. Dukarm and M. Duval. Transformer reliability and dissolved-gas analysis. In *2016 CIGRE Canada Conference*, number 807, Vancouver BC, October 2016.

[2] James J. Dukarm. Progress in transformer dissolved-gas analysis. *NETA World Journal*, Fall 2019.

[3] *Mineral oil-filled electrical equipment in service – Guidance on the interpretation of dissolved and free gases analysis*. Number IEC 60599-2015-09. International Electrotechnical Commission, 3.0 edition, Sep 2015.

[4] IEEE guide for the interpretation of gases generated in oil-immersed transformers. *IEEE Std C57.104-2008 (Revision of IEEE Std C57.104-1991)*, pages 1–36, Feb 2009.

[5] IEEE guide for the interpretation of gases generated in mineral oil-immersed transformers. *IEEE Std C57.104-2019 (Revision of IEEE Std C57.104-2008)*, pages 1–98, Nov 2019.

[6] JWG D1/A2.47. Technical Brochure no. 771 Advances in DGA interpretation. *CIGRE Technical Brochures*, 2019.

[7] M. Duval. Calculation of DGA limit values and sampling intervals in transformers in service, 2008.

[8] F. Jakob, P. Noble, and J. J. Dukarm. A thermodynamic approach to evaluation of the severity of transformer faults. *IEEE Transactions on Power Delivery*, 27(2):554–559, 2012.

[9] F. Jakob and J. J. Dukarm. Thermodynamic estimation of transformer fault severity. *IEEE Transactions on Power Delivery*, 30(4):1941–1948, 2015.

[10] L. Maxim, R. Niebo, and M. Utell. Screening tests: a review with examples. *Inhalation Toxicology*, 26:811–828, 2014.

[11] Afina S. Glas, Jeroen G. Lijmer, Martin H. Prins, Gouke J. Bonsel, and Patrick M. M. Bossuyt. The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology*, 56(11):1129–1135, 2003.

[12] Tsung-Hsien Chiang, Yi-Chia Lee, Chia-Hung Tu, Han-Mo Chiu, and Ming-Shiang Wu. Performance of the immunochemical fecal occult blood test in predicting lesions in the lower gastrointestinal tract. *CMAJ*, 183(13):1474–1481, 2011.

[13] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ROC Analysis in Pattern Recognition.

[14] Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 233–240, New York, NY, USA, 2006. Association for Computing Machinery.