We could be looking at an index to help plan for transformer replacement or to identify those which will be most in need of maintenance or refurbishment: identifying a need to act, an 'intervention'

Transformer health indices and risk matrices

Introduction

A transformer health index is an estimate of the health of a transformer: available data is analyzed, combined and 'boiled down' to a number, code or some value that can be used in planning or other asset management activities. The index does not tell us anything new about the transformer – we would have the same level of knowledge about the transformer if we did not generate an index. What is novel about an index is the ability to rank assets based on the index.

In this discussion, we will look at some of the key points to remember and / or address in generating and using an index; some of them are 'obvious,' but some are more subtle. However, it is critically important to start with the endpoint in mind so we can not only realize the value of the index but also see the limitations [1]. Risk matrices often use the output of a health index analysis to act as a proxy for 'probability of failure'; we will address some of the issues in this approach and some of the issues related to the categorization of both indices and risk matrices which can be misleading.

Setting objectives

The first question is: what problem are we trying to address [1]? We could be looking at an index to help plan for transformer replacement or to identify those which will be most in need of maintenance or refurbishment: identifying a need to act, an 'intervention.' If we think of a more familiar analogy, say a car, we might be interested in when it will need maintenance and when it will need replacement: but do we replace the car because the tires are going down in pressure? Probably not – we can just deal with the tires. They can be considered a maintainable item – even if maintenance, in this case, means replacement.

So: what is to be included in the transformer health index? Cooling? Oil containment? Bushings? Tap changer? Oil? Windings? You get to choose, making sure that whatever is included is used to

address the objective of the index [2].

An index is going to be generated by analysis of available data, possibly including numerical data such as dissolved gas analysis (DGA) values, winding resistance measurements, and so on, and more graphical data such as sweep frequency response charts (SFRA), phase resolved partial discharge (PRPD) plots and so on. In addition, we may be applying applicable standards and guidelines to interpret data and derive diagnoses, say of DGA data where several diagnostic approaches exist [3,4]. There are few standards or guides for analyzing SFRA data, and fewer still which quantify the results into an indication of transformer 'fitness for function'. Generating an index provides a means to bring the raw and derived data together – but how do we do so in a sensible manner, reflecting the diagnoses and prognoses of that data?

Suffice to say that we have an index, which helps with the objective, but we need to be aware of the range and quality of data that helps create that index: what the squiggles of SFRA mean and what the DGA levels imply.

And it may be that we need several different indices to cover different components and different applications: a bushing health index, a tap changer health index, a transformer health index and so on, and we could have separate indices for both maintenance and replacement. In each case, the index needs an objective and clarity of analysis. We could go further and have a DGA index, a PD index, and an offline testing index: we get to decide if that will help meet our objectives.

One question that is often missed is: how can we demonstrate that even though the index addresses the objective of the analysis, does it improve on other approaches, including the 'null' case of doing nothing or using a placebo such as 'replacement at random'? This is a good question as there are many possible index calculation methods, and we need to see the benefit of using a particular approach.

One question that is often missed is: how can we demonstrate that the index addresses the objective of the analysis, but does it improve on other approaches

APOLLO 11 Moon landing: go or no go?

Several times during the Apollo 11 Moon landing mission the flight controller can be heard asking individual members of the team their 'status': 'go' or 'no go.' As the controller goes round the team individually, they respond. Any one person could have caused the mission to abort at any point – but the go-round is a confirmation that there is nothing overtly wrong at that point. Similarly, the health index does, to a degree, the same thing: a good analysis will show that there is nothing overtly wrong. It cannot guarantee that the transformer will even last another 24 hours: we can only say that we see no reason from available data and analyses why it should not.

Expectations of a health index

Noting that a health index tells us nothing new, nothing we do not already know in terms of data and analyses, we should not expect it to tell us what needs to be done or when. That is the outcome of the engineering analysis of the data. This means that we do not perform an intervention because the health index has a particular value or because the health index 'says so.' Quite the reverse, the health index has a particular value as we have analyzed data, and as a result, we need to perform an intervention within a particular timescale. The difference is not subtle but is often lost on the more asset management types who understand numbers and spreadsheets, but are less familiar with SFRA, say, or leakage reactance (LR) results [5].

Health index as a digital twin?

We could call the health index a low-resolution digital twin: it summarizes what we know and what may happen over time with relation to the objective of the index. It is a ***model*** of the actual health of the transformer – it is NOT the actual health of the transformer. And, to quote a renowned statistician: "All models are wrong, some models are useful" (6). We do need to check that the index we use is proving to be useful. Heywood *et al.* compared predicted transformer health before planned replacement with that found when detailed tear down of the transformer took place (7). In many cases the prediction was very close to what was actually found, while in others there were discrepancies as a result of things which could not have been accounted for ahead of time: undocumented changes in design and materials for example. The result underscores the need to understand the precision of an index when it is used.

A health index may be given as a percentage score or may be given as a more 'digital' value: say, an integer value from 1 to 5 and we get to choose whether 1 is 'good' and 5 is 'bad' or vice versa

To address the index objective, will we need to be analogue or digital? Some index systems compress the data and produce a percentage value, which is meant to represent the health of the transformer overall on a scale of 1 to 100: an analogue value. Weighted systems often do this. But transformers do not, in general, fail for 'overall' reasons: they fail for a specific reason: there is a failure mode. Does the index reflect the failure modes in operation and the timescales which apply? If not, how can we expect to use the result as a replacement index? A 'digital' approach uses the available data and then puts the transformer in a category: the labels for the categories could be: 'poor', 'good' and so on, or they could relate to predicted timescales for intervention. The result is 'digital' as there are a limited number of categories to choose from.

Features of a health index

A health index may be given as a percentage score or may be given as a more 'digital' value: say, an integer value from 1 to 5. We get to choose whether 1 is 'good' and 5 is 'bad' or vice versa: an index of 1 may mean that the transformer should last more than 10 years, while a 5 means it needs to be replaced within 2 years – whatever helps meet the objectives of the index. The different index values are really just labels for health categories: we could use A through E, or labels such as 'good', 'poor', 'very poor', 'bad' and 'disaster', even though the words are imprecise themselves, they should relate to a well-defined set of criteria. (It is unlikely that anyone would wish to use the label 'disaster' in practice as it is somewhat emotive.) How the data is translated from the raw and derived values into a category should be clear: the trail from data to index should be both *auditable* and *justifiable*; the analyses and interpretations applied should make sense, both in terms of practical engineering and in terms of addressing the objective.

In addition, the index should relate to time: individual transformers may require different maintenance activities, have different failure modes in operation leading to replacement needs, be subject to different operational constraints, but the one thing which unites all the different

information is that we are looking to identify those which are more urgent and those which are less urgent. Consequently, an index needs to have two key features, those being:

- *Calibration*: the same index always implies the same timescale for intervention
- *Monotonicity*: that is, a worse score ALWAYS means a more urgent intervention is needed.

We get to choose timescales and index categories, but if the index does not have these two key features, calibration and monotonicity, we have little chance of justifying intervention and spending real money [8].

To achieve calibration and monotonicity, we need to identify both deterioration in progress and likely failure modes in operation. These can then be used to estimate timescales before intervention is needed, and, if appropriate, an estimate of the *probability of failure* (PoF) of the transformer.

Probabilities

Probability of failure (PoF) – working forward from data

As an analogy, think in terms of a car tire – of which there are billions around the world: what is the PoF in the next six months if we are at normal operating pressure? What if the pressure drops by 5 psi (~35k Pa)? How accurate is that probability: + / -? What if the pressure continues to drop – when does the PoF become unacceptable? How accurately can we calculate the tire probability of failure in a meaningful way? Not very well… We need to know the manufacturer, the road conditions, the driver's way of driving, possible deterioration mechanisms and so on. Noting there are way more tires than transformers, we must be cognizant of the fact that we will not get a precise PoF either for a tire or for a transformer. Working forward from raw data and derived data to generate a PoF is a very imprecise matter: experience and engineering judgment may help.

What is the probability that a transformer will fail? 100 %. "Everything put together sooner or later falls apart [9]" The real question is how likely a failure in a given timescale is: say over the next week, month, year, decade, and then whether it will last that long with recommended maintenance or some variation thereof.

Statistics

Statistics are values derived from a set of individual data points; statistics apply to the population, not the individual. Think of a large set of standard, fair, six-sided dice. If we roll all of them, we'd expect about 1 in 6 rolls to show a 1, with similar expectations for each other possible outcome. The 'expected value' of a roll is the average of all possible outcomes and can be calculated at 3.5. Problem is that expected value will never be rolled in practice. Thus: statistics apply to the population, not the individual. (Unless, of course, we are talking about an individual population, but that is just a tad too pedantic.)

We can also work backwards from known failure rates to indicate how many failures for a population we expect and then work out which are the worst performers and thus most likely to fail and thus fit the worst category of the index. This requires the data analyses which produce the indices to be monotonic and calibrated: but will next year look like last year or even previous years? And what if you only have one transformer? How do the statistics help then? Maybe, if we have a population of similar transformers, as described by Doble's transformer failure analyses [10, 11].

Translating data into an index

First rule: never forget the index represents our *interpretation* of the raw data. It may be in a convenient form for analysis, but it is a very low-resolution version of the original data.

We can generate an index based on whatever data is available: it may be that the manufacturer is a good indicator of the reliability of a particular transformer type and can be used to provide an index of low precision using some simple categories and / or timescales.

For each piece of data available: does it indicate a possible problem related to the index objective? (Reference back to the Apollo 11 box). If it does, what is the urgency? If we use the data in conjunction with other pieces of data, we may come to the same conclusion as to urgency, but now we would be more confident in the precision of the answer.

Never forget the index represents our *interpretation* of the raw data, it may be in a convenient form for analysis, but it is a very low-resolution version of the original data

An index example: let's look at a familiar scenario and address whether a tire is in good condition or poor condition, based on the actual pressure of the tire. For the sake of the example, the expected tire pressure nameplate value is 34 psi, and if the pressure drops below 30, we should address it within three months. So, we have two indices, A and B, as per Table 1, where we have also given a condition label (or category) and a numeric code of 1 for A and 2 for B.

If we look at a particular set of four tires, we could find that maybe 3 are code A and 1 is a code B. We'd know that we had to do something within three months – which we knew anyway from the raw data. We'd get the same message from the labels and the alpha / numeric codes.

The interesting thing is we know what we need to do and when – but with numeric codes, there is a temptation to start to do the math on the values: the average for our particular set of tires is 1.25. But what does that mean? True, it represents the mean of the population but no longer represents any individual and does not represent the urgency of the situation.

The same can be said to apply if we don't have 4 tires but 4 dissimilar components. Let us, this time, have a set of five codes with alphanumeric and labels equivalents; this time, the intervention may well be different for each component, but the timescales need to be consistent for the monotonicity and calibration requirements, as per Table 2.

With numeric codes for the health index, we know what we need to do and when, but there is a temptation to start to do the math on the values whose interpretation may be questionable

Table 1. Simple index labels and alphanumeric equivalents

Index category label	Description	Alpha code	Numeric code
Good	Tire should be ok for >3 months	A	1
Bad	Tire needs attention in <3 months	B	2

Table 2. Five health index codes with alphanumeric equivalents

Index category label	Description	Alpha code	Numeric code
Good	Intervention in >10 years	A	1
Poor	Intervention in >5 & <10 years	B	2
Very poor	Intervention in >2 & <5 years	C	3
Bad	Intervention in >1 & <2 years	D	4
Disaster	Intervention in >0 & < 1 year	E	5

Table 3. Components rated on calibrated / monotonic indices

Component	Label	Alpha code	Numeric code
Bushings	Very poor	C	3
Tap changer	Good	A	1
Windings	Good	A	1
Oil	Bad	D	4
Cooling systems	Poor	B	2

We can do lots of interesting math, applying weights as multipliers before averaging and other mathematical tricks: but we do need to be aware of the practicality of the analysis

And let us say the components have been individually reviewed based on their data, and planned interventions and timescales are appropriate for their condition: for bushings, tap changers, windings, oil and cooling systems. A particular set of codes could be as shown in Table 3.

The urgency of the situation is identified by looking at the worst component: whether via label, alpha code or numeric code. The oil is the driver for intervention in Table 3, being put in the 'Bad' category, otherwise labelled 'D' or '4'. We can still do math on the numeric values and get a result for the small population of components – the average being 2.2. But, again, the average is pretty meaningless: it doesn't retain the urgency and gives no indication of the spread of results: in fact, the same average could be from components with scores 1, 1, 1, 3, 5, which, in fact, would be more urgent as it has a code 5 in the data.

We can do lots of interesting math, applying weights as multipliers before averaging, and other mathematical tricks: but we do need to be aware of the practicality of the analysis – a theme discussed by Dr. Hall in her work on "Tooth Fairy Science" [12], where she notes that performing analyses which are reproducible, consistent and statistically significant doesn't mean we learn anything about what we are interested in. The article is illuminating and well worth reading: how do we check that what we are doing addresses the practical question in a meaningful manner?

A more natural approach is to look for the maximum urgency: this can be derived from the original data. In fact, by looking only at the resulting labels / categories, we lose the precision of the original data: hence the request from some organizations to help them prioritize their transformers once they have generated the health index for each – say, all the units with a maximum code of 3. One issue is the effect of using the labels/categories at all: as discussed by Prof. Sapolsky of Stanford University [13]. Thinking in categories may, in fact, be misleading because as soon as we use categories, we tend to do two things:

- underestimate how different two things may be when they fall in the same category,
- overestimate how different two things may be when there is a boundary between them.

Prof. Sapolsky's lectures are well worth reviewing as they give clear examples of the issues faced, which we will discuss in greater detail in the next section.

One thing we could do with just category data, without mathematical trickery, would be to look at the raw data which made the component a code 'Bad' or 'D' or 4, say, and use that as a guide for urgency. This is what we would do if we didn't compress all the data into an index – say, if we just had one transformer to deal with. We should note there is a precision relating to the raw data, so the data itself could put us anywhere in a 1–2-year timescale for range. But, as Prof. Sapolsky notes, we may well be near a boundary, and a small change in the data could push us across that boundary. This information is lost when we look only at categories and not the raw data.

There is, however, an intermediate step that can be taken, using just the labels/categories in the table. The approach is called 'Enumeration' and is extended here to three new transformers in Table 4, where

Table 4. Codes for three transformers

Transformer	T1		T2		T3	
Component	α	N	α	N	α	N
Bushings	C	3	A	1	B	2
Tap changer	B	2	A	1	B	2
Windings	A	1	A	1	A	1
Oil	D	4	D	4	E	5
Cooling systems	B	2	E	5	B	2

Table 5. Enumeration approach

Transformer	Code E/5	Code D/4	Code C/3	Code B/2	Code A/1	Enum	Urgency
T1	0	1	1	2	1	01121	3rd
T2	1	1	0	0	3	11003	1st
T3	1	0	0	3	1	10031	2nd

we just use two codes: α is the Alpha code, and N is the numeric code.

Of the three transformers in Table 4, T2 and T3 both have a component with the highest urgency code of E or 5; but the oil in T3 is a higher code, so maybe that puts the whole transformer in an overall worse condition. It would be best to see just where in the category the data lies – closer to one year or closer to immediately – but a simple indication is found by looking at how many code 5's there are, then how many code 4's, 3's etc. and using the resulting number to rank the transformers, as shown in Table 5 by starting with the highest enumeration. (Note the average for each transformer in Table 4 is 2.4: which makes it difficult to prioritize based on averaging, which is a linear uniform weighting; we could try to weight things non-uniformly, but that would highlight some components over others, and failure tends to be for specific rather than general reasons).

Note that the enumeration approach is not a replacement for looking at the raw data, as enumeration assumes each categorization is identical in urgency – which is not necessarily the case. The issues relating to categories are presented by Prof. Sapolsky but are also discussed in terms of risk matrices by Bratvold *et al.* [14], referenced in the next section.

Enumeration approach is not a replacement for looking at the raw data, as enumeration assumes each categorization is identical in urgency – which is not necessarily the case

Risk and index issues

The health index is driven by technical data and interpretation; whether to act on the data and the interpretation is a business decision: a risk management activity.

From ISO 31000, the risk is the "effect of uncertainty on objectives" [15]. This sounds simple enough, but what does it really mean? It means we need to know what our 'objectives' are and the consequence of things that may happen, which mean we may not meet the objectives. If my objective is to drive my car safely and not be interrupted by needing to change a tire, I may wish to address any issues with the tires sooner rather than later: but that doesn't address the possibility that I drive over a sharp nail, lying in the road, and burst a tire… That failure mode is difficult to address through prevention. If, however, I think the tire will fail in a way that is 'benign', I may choose to let it fail and then use the spare to replace it and carry on. Always assuming I have a spare and can make it happen. Or I can call the road service team: AAA in the USA, say, to fix things up and get me going. It all comes down to what is the 'hazard' which we need to address, and a 'nail in the road' is a different hazard to 'worn tires' in terms of both timescale to failure and in terms of possible consequences. Note 'risk' is not just a 'probability'. It is a combination of a hazard and the probability the hazard may occur, or be 'realized', and the consequence of that occurrence. If we know the probability of a hazard occurring in a given year, say, and the consequence in a monetary equivalent, then we multiply the two together to give our risk magnitude in money / year.

In general, we need to look at mitigating the risk by either reducing the likelihood of the hazard occurring or reducing the consequence of it occurring. But we must never lose sight of the original data: the likelihood and consequence, as some risks, maybe a very low probability but have a significant consequence, while others are much more likely with a lower consequence: the resulting risk magnitude may be similar, but the strategies to deal with them are likely to be very different as described by Buckland [16] and summarized in Fig. 1.

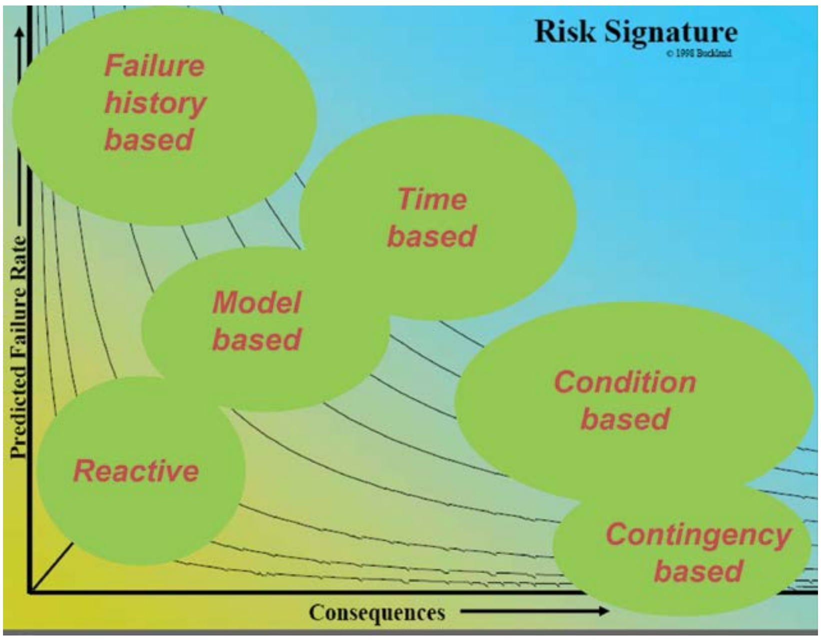

Figure 1. Risk strategy chart (from Buckland)

When managing a range of transformers, it is not uncommon to use a risk matrix or risk heat map to plot the position of all transformers on a single chart

Buckland identifies different strategies for different areas of the risk chart, Fig. 1, depending on the probability and consequence of the hazard. Those strategies will rely on knowledge of the individual assets and components and their failure modes, failure rates and probably consequences. For some cases, for example, a reactive strategy is appropriate. However, for larger consequences, a more pro-active approach is recommended, based on condition and operational contingencies.

When managing a range of transformers, it is not uncommon to use a risk matrix or risk heat map to plot the position of all transformers on a single chart. The result may be colour coded green-yellow-red to indicate increasing levels of risk, as in Fig. 2, where a number of individual transformers have been placed on the chart.

The approach may be familiar – it is fairly easy to understand, and the visuals make things quite comprehensible: green = good, red = bad, higher and to the right is worse. But there are some issues.

As Bratvold notes, the use of risk matrices has not been shown to be more effective as a risk management tool than other commonly available tools; worse, the risk matrices themselves can be quite misleading, with several different issues spelt out:

- Range compression: same category for very different risks
- Centring bias: people avoid the extremes when asked to estimate consequence / probability magnitudes
- Category definition bias: people make the definitions fit their experience
- Ranking reversal issues: changing scales can re-order risk ranking
- Lie factor: the distance between transformers on the chart does not always represent a real difference in practice
- Use of 'very low', 'low', 'medium', 'high', 'very high' can be misleading in interpretation and lead to very poor decision making (much more detailed discussion can be found in Hubbard [17]).

The paper from Bratvold was presented at an oil industry event, but the lessons apply equally to transformers and other assets, and the pitfalls of categories are valid. To look at just one area, the 'lie factor' consider the fact from Sapolsky that we overestimate the similarity of items within a category and underestimate the similarity of items in different categories: transformers 11 and 9 may seem far apart but may, in fact, both be on the L/M boundary and quite close together when we consider their raw data and the risk magnitude, while transformers 2 and 6 may be at opposite edges of the two categories and very different in their raw data and thus very different in risk magnitude.

Conclusions

Even though health indexing is common, there is often little attention paid to the details of the index derivation and its relation to the raw data and urgency of intervention, and there may be no benchmarking of index systems to show the aim of the index is actually met in practice. In this discussion, we have shown some of the pitfalls of index derivation and some of the issues relating to the use of categories in presenting data. But there are still index systems in place which feed into asset management tools where decisions are being made based on risk matrices not dissimilar to those shown above. The result is that people with an ability to understand numbers and spreadsheets will likely have overconfidence in their ability to make decisions, based on implications of the Dunning-Kruger effect where 'people with a little knowledge usually have overconfidence in their ability' [18] which has been found to apply across many industries. What we need is a better understanding of the raw and derived technical data, the failure modes in operation, the timescales

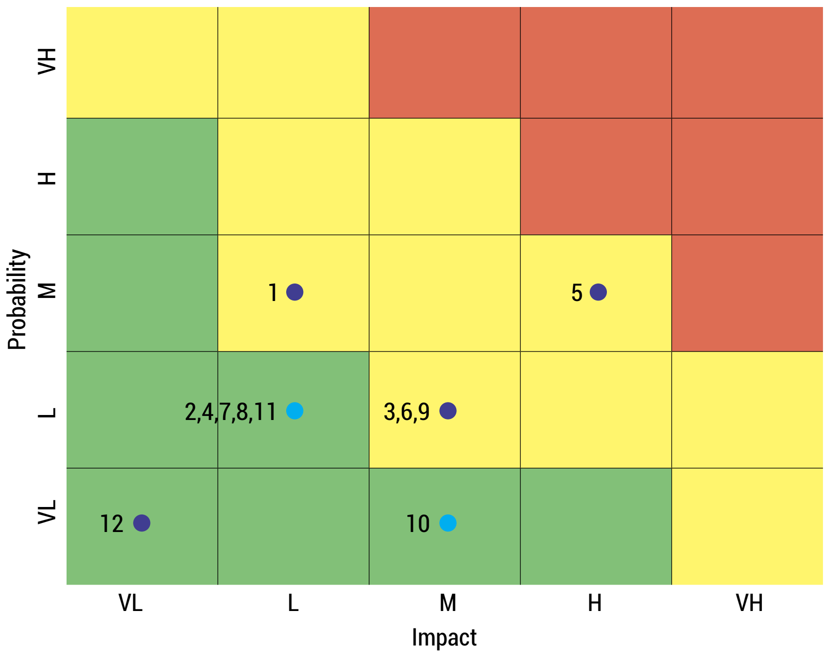

Figure 2. Risk matrix chart (from Bratvold)

Even though health indexing is common, there is often little attention paid to the details of the index derivation and its relation to the raw data and the urgency of intervention

involved, the imprecision on both timescales and PoF, and the asset management context for technical decisions. The technical data tells us what we need to do and by when we should do it; the asset management context helps us decide whether to do it sooner, later or at all.

Key points relating to health indices

1. Why bother? We need clarity of objective – otherwise how do you know you have prioritized correctly?
2. Consider the health index as a very low-resolution digital twin.
3. Focus: Could several indices be simultaneously in use if they are useful?
4. Which components are included: each could have its own index, too?
5. Subcomponents could also have indices if that is useful / valuable.
6. The index should be:
   a. Calibrated – for time, hence urgency, to make planning sense.
   b. Monotonic – allows for sensible ranking based on urgency.
   c. Auditable – can show a path from raw data through analytics to final index.
   d. Justifiable – the path to the index is based on engineering and analyses, not magic.
   e. Benchmarked – have we chosen the best system for our application? Or just the first one we saw?
   f. Able to identify failure modes – and where we are on the path to failure.
7. Precision: the data / interpretations are imprecise, so we need to understand the effect on the final index.
8. Probability of failure: working forward from data to failure is both difficult and very imprecise.
9. Probability of failure: using historic rates may be misleading as the population and operational regimes / contingencies vary.
10. Thinking in categories leads us away from raw data and promotes poor thinking.
11. Close the loop – check the health index with what is actually in the transformer through internal inspections and tear downs.
12. Re-evaluate the validity and value of the index system regularly: is it still useful?

Acknowledgements

The author wishes to thank the many people who have, over several years, discussed, dissected and analyzed the concepts and ideas presented here.

Bibliography

[1] S. Covey, *The 7 Habits of highly effective people*, Franklin-Covey, 2004

[2] R. Heywood, T. McGrail, *Clarifying the link between data, diagnosis and asset health indices*, Institute of Asset Management Annual Conference, London, UK, 2015

[3] *IEEE Guide for the interpretation of gases generated in mineral oil transformers*, IEEE C57.104-2019

[4] *Condition assessment of power transformers*, CIGRE Technical Brochure 761, 2019

[5] M. Wolf, *50 MVA transformer failure as experienced by maintenance engineers*, 79th International Conference of Doble Clients, Boston, MA, 2012

[6] C. J. Box, *Science and statistics*, Journal of the American Statistical association, #71, 1976

[7] Cross E. et al., *Learning from power transformer forensic investigation and failure analysis*, CIGRE Conference, Paris, France, 2014

[8] T. McGrail, *Think of a Number: then spend millions on it*, Institute of Asset Management N. American Conference, Chicago, USA, 2019

[9] Song Lyric from *Everything put together falls apart*, Paul Simon, 1972

[10] R. Hernandez, M. Rivers, *Lessons learned from analysis of historical power transformer failures*, 86th International Conference of Doble Clients, Boston, MA, 2019

[11] R. Hernandez, M. Lachman, *Refining bushing power factor and capacitance analysis through statistics*, Transformers Magazine, Vol. 7 Issue 3

[12] https://skepticalinquirer.org/exclusive/tooth-fairy-science-part-1/

[13] Stanford University, *Introduction to human behavioral biology*, https://www.youtube.com/watch?v=NNnIGh9g6fA

[14] P. Thomas, R. Bratvold, J. Bickel, *The risk of using risk matrices*, Soc. Petroleum Engineers, Annual Tech Conf., New Orleans, USA, 2013

[15] ISO 31000, *Risk management*, 2018

[16] P. Buckland, *Asset management: The right road for the long haul*, ICOMS 2005 Conference, Australia

[17] D. Hubbard., *How to measure anything*, 3rd Edition, Wiley, 2014

[18] J. Kruger, D. Dunning, *Unskilled and unaware of it*, Journal of Personality and Social Psychology #77, 1999

.

Author

Dr. Tony McGrail of Doble Engineering Company provides condition, criticality, and risk analysis for substation owner / operators. Previously, he has spent over 10 years with National Grid in the UK and the US as a substation equipment specialist, with a focus on power transformers, circuit breakers, and integrated condition monitoring, and has also taken on the role of substation asset manager identifying risks and opportunities for investment in ageing infrastructure. McGrail is a Fellow of the IET, past-Chairman of the IET Council, a member of the IEEE, ASTM, ISO, CIGRÉ, and the IAM, and is a contributor to SFRA and other standards.