Modelling and Simulation of Detection Rates of Emergent Behaviours in System Integration Test Regimes

Rune André Haugen[1,2] Ali Ghaderi[3]

[1]Department of Science and Industry Systems, University of South-Eastern Norway, rune.a.haugen@usn.no
[2]Kongsberg Defence and Aerospace AS, Norway, rune.andre.haugen@kongsberg.com
[3]Department of Mathematics and Science Education, University of South-Eastern Norway, ali.ghaderi@usn.no

Abstract

System level testing generally lacks coverage due to cost of performing realistic tests on the "system as a whole". This lack in test coverage gives rise to seemingly emergent behaviour at system level. The interactions between multiple sub-systems lead to "the whole being greater than the sum of its parts", which is a famous saying dated back to the time of the Greek philosopher Aristotle. Either we should test more extensively at system level, or we should test smarter. The company needs to validate its current test regime to see if the current way of testing detects the emergent behaviours in question. We seek to validate the company's system integration test regime to see if it can detect a given set of emergent behaviours. This paper aims to find the probabilities of detecting specified types of emergent behaviour in the way the company performs system integration testing today and compare that to alternative test regimes. A model is set-up to find the probabilities of the emergent behaviour types in the different test regimes, and to simulate the corresponding detection rates and related uncertainties. The results show that the company could benefit from changing to an alternative test regime, which has higher probability of detecting a given set of unwanted behaviours emerging through system integration testing.

Keywords: *Bayes' theorem, emergent behaviour, experimental design, statistical inference, system integration testing.*

1 Introduction

System level testing generally lacks coverage due to cost of performing realistic tests on the "system as a whole". This lack in test coverage gives rise to seemingly emergent behaviour at system level. The interactions between multiple sub-systems lead to "the whole being greater than the sum of its parts", which is a famous saying dated back to the time of the Greek philosopher Aristotle. Either we should test more extensively at system level, or we should test smarter. The company needs to validate its current test regime to see if the current way of testing detects the emergent behaviours in question.

This paper looks at how well the system-level test regime detects unwanted behaviours for an autonomous underwater vehicle (AUV) that uses a camera to capture images of the current seabed.

We seek to validate the company's system integration test regime's ability to detect a given set of emergent behaviours. This paper aims to find the probabilities of detecting specified types of emergent behaviour in the way the company performs system integration testing today and compare that to alternative test regimes.

1.1 Problem Statement

The company performs system integration testing based on manual operations, which is a bottleneck for them to ensure mature and robust products (Haugen and Mansouri, 2020).

Analysts in the company do not have enough time to analyse all available test results from performed test executions / simulations. Roughly, system domain experts analyse 10% of test results on average. About 80% of tests analysed contain no errors. Around 20% of tests with errors detected include behaviour-related errors (Kjeldaas et al., 2021). Illustration in Figure 1.

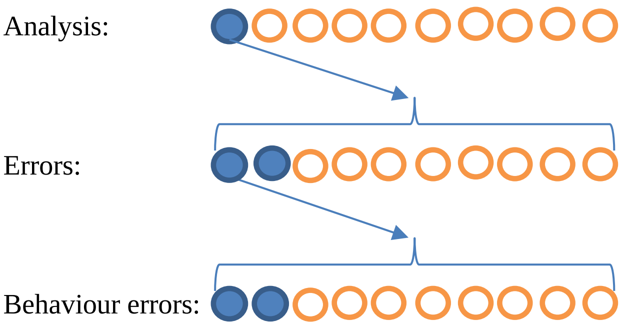

Figure 1. Portion of tests with detected behaviour related errors.

We believe the company tests too many "sunny day" scenarios compared to "rainy day" scenarios. This test strategy fails to trigger the system's inherent emergent behaviours to the extent that we can collect enough data on them through testing to perform effective analyses of these behaviour issues.

The AUV uses available map data to plan missions. The map data varies in quality, which may give problems for the accuracy of the planning functionality. Map areas lacking data works as tripwires for the planning system and could cause the planning to fail if it is not possible to avoid these areas, ref. Figure 2.

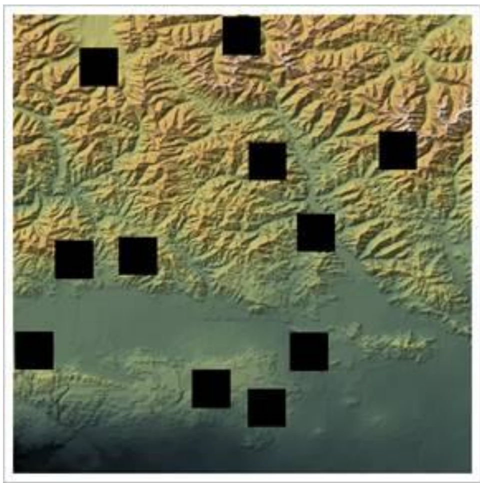

Figure 2. Principle sketch of issue with lacking (black areas) map data.

The AUV plans route segments within available fuel limit, including departure from -and arrival to the mothership. Complex ocean currents yield large fuel calculation error margins. Reaching fuel point of no return forces the AUV to abandon mission and return to mothership, ref. Figure 3.

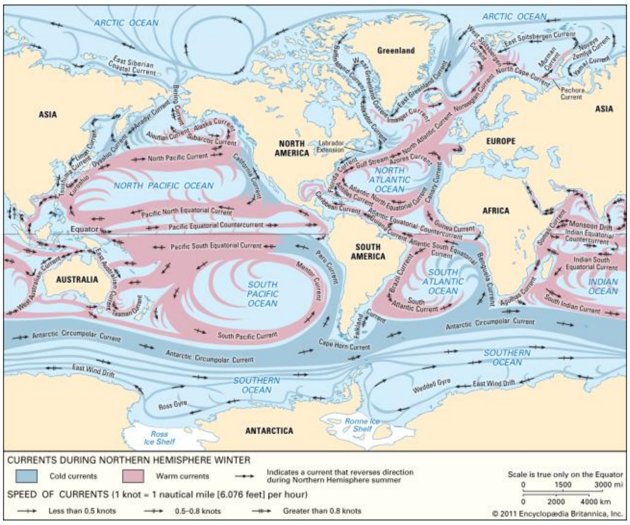

Figure 3. Principle sketch of issue with complex ocean currents (Cenedese and Gordon, 2021).

The height information available of the seabed have varying uncertainty, which is a critical factor for the AUV's ability to capture seabed images of sufficient quality. To ensure desired quality in the photograph of a given area, the AUV needs a minimum number of pictures of the same area. If the slope of a ridge is too steep, the AUV does not have time to photograph the slope with sufficient quality, or photograph it at all, without special considerations in planning the route, ref. Figure 4.

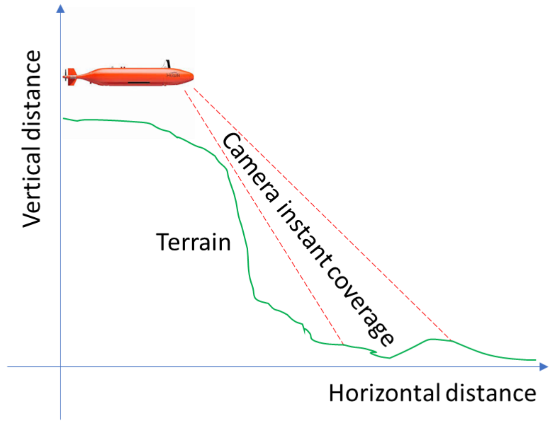

Figure 4. Principle sketch of issue with steep ridges.

The AUV uses an acoustic positioning system (APS) to keep on track with the planned route. If the AUV APS information is lost due to some interferences, the AUV drifts from its planned route depending on the inertial navigation system (INS) and terrain correlation, ref. Figure 5.

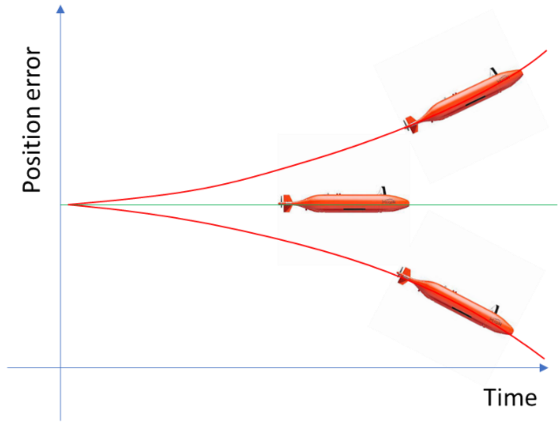

Figure 5. Principle sketch of issue with navigation drift.

The company assumes the AUV system is complicated and even complex. Complex systems are understood only in retrospect and do not usually repeat, while complicated systems can be understood by reductionism and detailed analysis. The company assumes that the AUV system exhibits weak emergence, and potentially strong emergence. Strong emergence is unpredictable and inconsistent in simulations, while weak emergence is predictable and consistently reproducible in simulations (Mittal et al., 2018).

1.2 Methods

The company uses the *Changing One Single Thing at a time* (COST) or *Only one Factor At a Time* (OFAT) model (Montgomery, 2017). We use the COST/OFAT principle for the first test regime in this paper.

For the second test regime, we use a *two-level full factorial design* (Dunn, 2021) and (Montgomery, 2017).

For the third test regime, we use an optimum design for maximizing the probability of detecting the emergent behaviours in question.
We seek to answer the following research questions:

- How well is the company able to detect a given set of emergent behaviours?
- What is the probability of the company detecting a given set of emergent behaviours in the current company test regime?
- How much can the company increase the detection of a given set of emergent behaviours in an alternative test regime?

1.3 Literature Review

The *Only one Factor At a Time* (OFAT) method consists of selecting a starting point, or baseline set of levels, for each factor, and then successively varying each factor over its range with the other factors held constant at the baseline level (Montgomery, 2017).

For a *two-level full factorial design*, we run the complete set of 2^k experiments, where k is the number of factors and 2 is the number of levels for each factor. The results of the experiments we use to quantify the importance of each factor. Indeed, for this purpose, linear regression models, considering both the single factor and two-factor effects are used in this paper. For example, in the case of two factor model, the following fitted regression model can be used to determine the importance of each factor (Dunn, 2021).

$$y = \beta_0 + \beta_A x_A + \beta_B x_B + \beta_{AB} x_{AB} \tag{1}$$

In the following, the total probability is calculated based on the inclusion-exclusion principle when the different emergent behaviours are independent but not disjoint events. We calculate the total probability by formulas regarding different number of factors (Allenby and Slomson, 2010):

$$P(A+B) = P(A) + P(B) - P(A,B) \tag{2}$$

$$\begin{aligned} P(A+B+C) = {} & P(A) + P(B) + P(C) \\ & - P(A,B) - P(A,C) \\ & - P(B,C) + P(A,B,C) \end{aligned} \tag{3}$$

$$\begin{aligned} & P(A_1 + A_2 \dots + A_n) \\ & = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P\left(A_i, A_j\right) \\ & + \sum_{i<j<k} P\left(A_i, A_j, A_k\right) + \cdots \\ & + (-1)^{n-1} \sum_{i<\cdots<n} P(A_1, A_2 \dots, A_n) \end{aligned} \tag{4}$$

The union of a four-factor probability is illustrated in Figure 6.

The Bayes' theorem is expressed as

$$P(X|Y,I) = \frac{P(Y|X,I)\, P(X|I)}{P(Y|I)} \tag{5}$$

where X is our hypothesis, Y is our data, and I is relevant available information. The various terms in Bayes' theorem have formal names. The quantity on the far right, $P(X|I)$, is called the prior probability; it represents our state of knowledge (or ignorance) about the truth of the hypothesis before we have analysed the current data. This is modified by the experimental measurements through the likelihood function, or $P(Y|X,I)$, and yields the posterior probability, $P(X|Y,I)$, representing our state of knowledge about the truth of the hypothesis in the light of the data. In a sense, Bayes' theorem encapsulates the process of learning. The denominator is often simply a normalization constant (not depending explicitly on the hypothesis). In some situations, like in model selection, this term plays a crucial role. For that reason, it is sometimes given the special name of evidence (Sivia and Skilling, 2006).

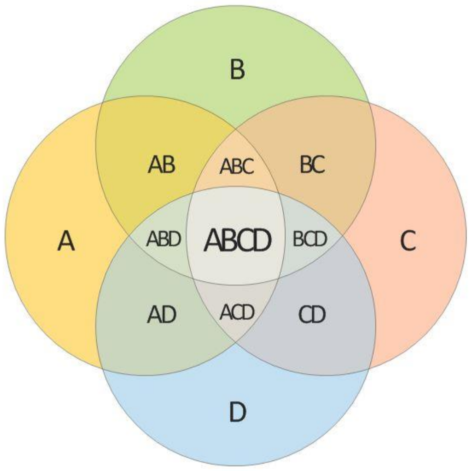

Figure 6. Inclusion-exclusion illustrated by a Venn diagram for four sets (Concept Draw, 2021).

2 Design of Experiment

This paper explores the probabilities and detections of a given set of emergent behaviours in different test regimes, the current company test regime and two other alternatives.

2.1 Data

We select a set of emergent behaviour types for study in this paper. The emergent behaviour types are:

- F1: Planning failure [0...1]
- F2: Fuel exceeded [0...1]
- F3: Photo quality degradation [0...1]
- F4: Photo coverage deviation [0...1]
- G: Any emergent behaviour [0…1]

For the purpose of this study, we have focused our efforts on a set of four dichotomous variables, which can take only two possible values (low and high). These are:

- A: Navigation quality [Low, High]
- B: Map delta height [Low, High]
- C: Real world environmental delta [Low, High]
- D: Map quality [Low, High]

Further, based on previous experience with comparable systems, we have selected a set of probabilities for this study. Accordingly, based on expert knowledge within the company, the probabilities of the emergent behaviours are assumed to be [%]:

- $P(F1) = 0.15$
- $P(F2) = 1.25$
- $P(F3) = 1.88$
- $P(F4) = 0.31$
- $P(G) = 3.59$

Moreover, based on the available data in the company's database, we will assume the following [%]:

- $P(D|F1) = 10$
- $P(D'|F1) = 90$
- $P(C|F2) = 55$
- $P(BD|F3) = 25$
- $P(BD'|F3) = 75$
- $P(A'|F4) = 100$
- $P(D) = 98$
- $P(D') = 2$
- $P(C) = 10$
- $P(BD) = 18$
- $P(BD') = 2$
- $P(A') = 1$

In general, we are interested in the probabilities for the different emergent behaviours at different factor levels. For example, we are interested in probability of planning failure (F1) under the condition that the map quality is high (D). That is, we are interested in $P(F1|D)$. This probability can be calculated using the Bayes' theorem (5),

$$P(F1|D) = \frac{P(D|F1)P(F1)}{P(D)} = \frac{10 * 0.15}{98} = 1.5 * 10^{-2}$$

The probabilities [%] of other emergent behaviours can similarly be calculated

- $P(F1|D') = 6.75$
- $P(F2|C) = 6.88$
- $P(F3|BD) = 2.6$
- $P(F3|BD') = 70.31$
- $P(F4|A') = 31.25$

2.2 Test Regime 1

The test regime 1 is the current company test regime and is the baseline for comparison with the other alternative test regimes. There are 16 possible combinations, using two values for 4 parameters. However, the company does not test all 16 cases. The principle is to start with a reference case and add cases with level-change in only one factor at a time as compared to the reference. This COST/OFAT principle makes it easier to analyse the effect of the level-change in one factor. Table 1 shows the company's selected test case types. Test case type 1 is the reference experiment type, and to reduce the number of test set-ups the company re-uses this as much as possible to verify system requirements. The company uses the test case types 2-5 to analyse the impact of the factors C, B, BD, and A, respectively.

Table 1. Scenario factor levels for test regime 1.

Test Case Type	A	B	C	D	# Runs
1	+	-	-	+	245
2	+	-	+	+	6
3	+	+	-	+	60
4	+	+	-	-	6
5	-	-	-	+	3

2.3 Test Regime 2

The test regime 2 is the first alternative test regime, which is also known as two-level full factorial experiment. The two-level full factorial design has four factors with two levels, which gives $2^4 = 16$ experiments. Table 2 shows the experiment set-up for test regime 2 with the yield for each test case type based on expert opinion that we choose for this study. The yield is the total number of emergent behaviour detections in 20 runs per test case type. Further, in the test regime 2, each test case is run equal number of times. For the purpose of comparison, we choose the total number of runs to be the same for both test regime 1 and 2.

Table 2. Scenario factor levels for test regime 2, full factorial design.

Test Case Type	A	B	C	D	Yield	# Runs
1	-	-	-	-	8	20
2	+	-	-	-	1	20
3	-	+	-	-	22	20
4	+	+	-	-	15	20
5	-	-	+	-	9	20
6	+	-	+	-	3	20
7	-	+	+	-	23	20
8	+	+	+	-	17	20
9	-	-	-	+	6	20
10	+	-	-	+	0	20
11	-	+	-	+	7	20
12	+	+	-	+	1	20
13	-	-	+	+	8	20
14	+	-	+	+	1	20
15	-	+	+	+	8	20
16	+	+	+	+	2	20

2.3.1 Effect of Experiment Factors

One way to find the effect of each factor on yield is by conducting a regression analysis based on the test results. In the case of the test regime 2, the Equation (1) has 16 parameters. The first coefficient being the average of all the yield values, while the other coefficients represent the effects of the different factors and factor interactions. Estimating the parameters with respect to the observed yield, results in the following relation with only six non-zero coefficients,

$$y = 8.14 - 3.13x_A + 3.65x_B + 0.69x_C - 4.06x_D - 3.39x_{BD}$$

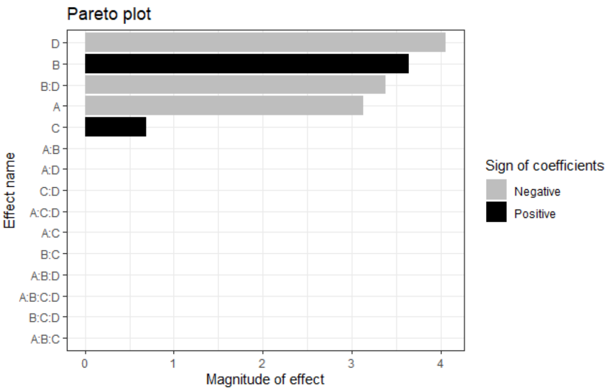

Figure 7. Magnitude of the effect of the factors.

The coefficient $-3.13x_A$ of factor A means that A at high level has a negative effect on the detection of the emergent behaviour. The coefficients are calculated for one step, but the regression model uses two steps from low to high. Therefore, the test regime 2 gives on average 6.25 more detections of any emergent behaviour type on a test case run 20 times with factor A at low level compared to high level. We can see the calculated factor coefficients in a Pareto plot (see Figure 7). Factor D has the highest impact on the detection of emergent behaviour among the main factors, while factor C has the lowest impact. The only active two-factor interaction is BD.

2.4 Test Regime 3

The test regime 3 is the second alternative test regime and is designed to optimize the detection of the given emergent behaviour types. The optimum way of detecting the emergent behaviours in question is to run the test case type(s) which have the highest probability of detecting the different emergent behaviour types. The probabilities for the different emergent behaviours were calculated based on Bayes' theorem in Section 2.1.

From Table 3 we see that test case type number 7 is the optimal test case type for triggering all emergent behaviour types. In the test regime 3, one runs only the case type number 7. However, the number of replicates is the same as the total number of tests run in other test regimes.

Table 3. Scenario factor levels and probabilities [%] for test regime 3, optimizing test design.

Test Case Type	A	B	C	D	F1	F2	F3	F4
1	-	-	-	-	6.75	0	0	31.25
2	+	-	-	-	6.75	0	0	0
3	-	+	-	-	6.75	0	70.31	31.25
4	+	+	-	-	6.75	0	70.31	0
5	-	-	+	-	6.75	6.88	0	31.25
6	+	-	+	-	6.75	6.88	0	0
7	-	+	+	-	6.75	6.88	70.31	31.25
8	+	+	+	-	6.75	6.88	70.31	0
9	-	-	-	+	0.015	0	0	31.25
10	+	-	-	+	0.015	0	0	0
11	-	+	-	+	0.015	0	2.6	31.25
12	+	+	-	+	0.015	0	2.6	0
13	-	-	+	+	0.015	6.88	0	31.25
14	+	-	+	+	0.015	6.88	0	0
15	-	+	+	+	0.015	6.88	2.6	31.25
16	+	+	+	+	0.015	6.88	2.6	0

3 Results

In this section the capability of the different test regimes in detecting any given emergent behaviour types is evaluated. The emergent behaviour type F1 has a single factor dependency in D. The formula for finding the probability of emergent behaviour type F1 in test regime 1, follows from the application of the marginalisation and product rule of the probability theory (Sivia and Skilling, 2006):

$$\begin{aligned} P(F1|T1) &= P(F1, D|T1) + P(F1, D'|T1) \\ &= P(F1|D, T1)P(D|T1) \\ &\quad + P(F1|D', T1)P(D'|T1) \end{aligned} \tag{6}$$

Further note that

$$P(D'|T1) = 1 - P(D|T1) \tag{7}$$

thus

$$\begin{aligned} P(F1|T1) &= P(F1|D', T1) \\ &\quad + \big(P(F1|D, T1)P(D|T1) \\ &\quad - P(F1|D', T1)\big)P(D|T1) \end{aligned} \tag{8}$$

and hence

$$\frac{P(F1|T1)}{P(F1|D',T1)} = 1 - \left(1 - \frac{P(F1|D,T1)}{P(F1|D',T1)}\right) P(D|T1) \quad (9)$$

Since the detection of the emergent behaviours depends only on the factors, then $P(F1|D',T1) = P(F1|D')$. The specific probabilities like $P(F1|D')$ are determined based on the abovementioned method in Section 2.1. We can then use a more general formula where we can separate the physical processes that we cannot control from the test set-up that we can control. The formula for finding the probability of emergent behaviour type F1 in test regime 1 is then:

$$\frac{P(F1|T1)}{P(F1|D')} = 1 - \left(1 - \frac{P(F1|D)}{P(F1|D')}\right) P(D|T1) \quad (10)$$

Using the results in Section 2.1 we get:

$$P(F1|T1) = \left(1 - \left(1 - \frac{0.015}{6.75}\right) * 98\right) * 6.75 = 0.15$$

Note that we have a generalized formula where we can replace T1 with T2 or T3. Indeed, on the right-hand side of the Equation (10), the choice of test regime only changes $P(D|T1)$. On the left-hand side, the denominator is fixed, which means that the change on the right-hand side can only affect $P(F1|T1)$. Consequently, we can then calculate the lower and upper bounds for detecting the emergent behaviour types by setting $P(D|T1) = 0$ and $P(D|T1) = 1$, ref. Table 5. The same principle applies to all emergent behaviour types with a single factor dependency (F1 and D, F2 and C, F4 and A) in the different test regimes (T1, T2, and T3).

The emergent behaviour type F3 has a two-factor dependency in BD. Although the final formula is different, it is also derived from the sum and product rule of the probabilities. Indeed, the formula for the probability of emergent behaviour type F3 in test regime 1 is:

$$P(F3|T1) = P(F3,BD|T1) + P(F3,BD'|T1) + P(F3,B'D|T1) + P(F3,B'D'|T1) \quad (11)$$

Further note that:

$$P(F3,BD|T1) = P(F3|BD,T1)P(BD|T1) \quad (12)$$

$$P(F3,BD'|T1) = P(F3|BD',T1)P(BD'|T1) \quad (13)$$

$$P(F3,B'D|T1) = P(F3|B'D,T1)P(B'D|T1) \quad (14)$$

$$P\left(F3,B'^{D'}|T1\right) = P\left(F3|B'^{D'},T1\right)P\left(B'^{D'}|T1\right) \quad (15)$$

thus

$$\frac{P(F3|T1)}{P(F3|BD',T1) + P(F3|B'D',T1)} = \frac{P(F3|BD,T1)}{P(F3|BD',T1) + P(F3|B'D',T1)} P(BD|T1) + \frac{P(F3|BD',T1)}{P(F3|BD',T1) + P(F3|B'D',T)} P(BD'|T1) + \frac{P(F3|B'D,T1)}{P(F3|BD',T1) + P(F3|B'D',T1)} P(B'D|T1) + \frac{P(F3|B'D',T1)}{P(F3|BD',T1) + P(F3|B'D',T1)} P(B'D'|T1) \quad (16)$$

Given the information:

$$P(F3|B'D,T1) = P(F3|B'D',T1) = P(B'D|T1) = P(B'D'|T1) = 0$$

Furthermore, using the results in Section 2.1 we get:

$$P(F3|T1) = \left(\frac{2.6}{70.31} * 18 + 2\right) * 70.31 = 1.87$$

The same principle applies to all other test regimes (T2, and T3). See Table 4 for the complete set of probabilities from the model.

Table 4. Calculated probabilities [%] for emergent behaviour types.

	TR1	TR2	TR3
F1	0.15	3.38	6.75
F2	0.69	3.44	6.88
F3	1.87	18.29	70.31
F4	0.31	15.62	31.25
G	3	36.84	89.17

We see from Table 4 and Table 5 that test regime 3 is at the upper bound and are the optimum way of testing to maximize detection of the emergent behaviour types. The total (G) is calculated using the inclusion-exclusion principle for the probability (Allenby and Slomson, 2010). The optimum test regime for detecting the given set of emergent behaviours has a probability of ~89% of detecting any given emergent behaviour, while the current test regime has only probability of ~3%. The test regime 3 can be used as the baseline in order to evaluate the capabilities of the other test regimes (see Table 5). Test regime 2 is detecting about half of the given emergent behaviours compared to Test Regime 3, while the test regime 1 is barely detecting any emergent behaviours at all.

Table 5. Calculated lower and upper bounds and relative frequencies for emergent behaviour types.

%	Lower bound	Upper bound	TR1	TR2	TR3
F1	0.15	6.75	2.22	50.1	100
F2	0	6.88	10	50	100
F3	0	70.31	2.67	26.01	100

%	Lower bound	Upper bound	TR1	TR2	TR3
F4	0	31.25	1	50	100
G	0.15	89.17	3.37	41.31	100

The probabilities in Table 4 can be used to answer many questions related to emergent behaviours. For example, if one chooses a test regime consisting of n runs, how many emergent failures of different types are expected to be detected? For each run, the probability of detecting a failure, say F1 in test regime 1, is $P(F1|T1)$. In each run, one either detects F1 or not. Moreover, since the runs are independent, then the probability of detecting k failures of type F1, follows a binomial distribution (n, p), for which $P = P(F1|T1)$. In Figure 8, we have simulated the probabilities for the detections of the different failures for the aforementioned three test regimes in 320 runs to find the detection rates of the emergent behaviours and the related uncertainties.

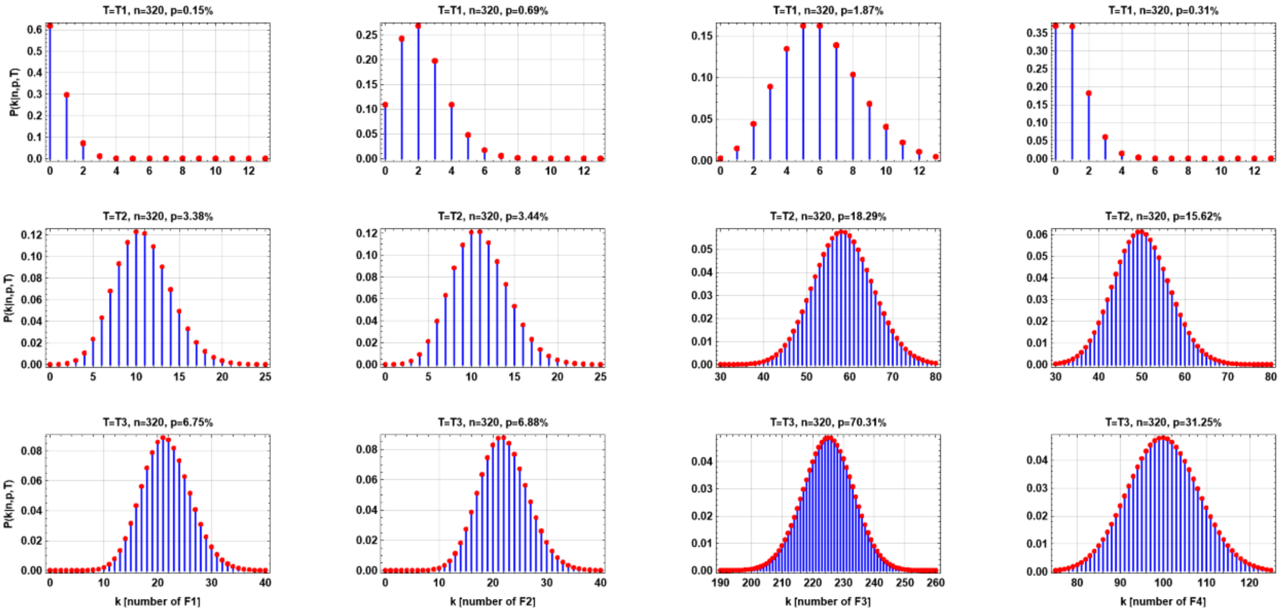

Figure 8. The probabilities of detections of emergent behaviours in different test regimes. Each row, from top to bottom, corresponds to a given test regime, T1, T2 and T3, respectively. Each column corresponds to a given failure type. In the present case the total number of simulated runs is 320.

4 Discussion

The company should increase the test analysis coverage at system level in their projects. The current test analysis coverage is in-sufficient to detect all emergent behaviour types of the system under test. The company cannot increase the test analysis coverage without automating the test result analysis. The test result analysis is the main bottleneck of the test system, and it is therefore crucial to make the analysis work more efficient.

For the company to stay competitive in the future underwater industry market they need to be able to run projects faster and run more projects in parallel. The automation of test result analysis is necessary to make the transition from the current test system to the desired future test system.

For the company not to have latent undesired emergent behaviour in their products, the test analysis needs to detect these with high enough probabilities. The test regime needs to change in the direction of triggering more of the emergent behaviour types of the system and trigger them with higher probabilities. The company will have better data to perform analysis of the emergent behaviours if the test regime triggers all emergent behaviour types of the system sufficient times in different scenarios. The company can get more insight into why the emergent behaviour types are triggered through deductive logic (Sivia and Skilling, 2006), and decide if they can do something to prevent or reduce the unwanted behaviours or the unwanted effects.

A combination of the different test regimes analysed in this paper may be the best approach for the company to deal with this problem of emergent behaviours. Test regime 3 triggers most emergent behaviours but does not see the effect of different settings. Test regime 3 satisfies the need to detect emergent behaviours by triggering emergent behaviours in about 89% of the tests. Test regime 2 sees the effect of different settings but does not trigger as much emergent behaviours as test regime 3. Test regime 2 also satisfies the need to detect emergent behaviours by triggering emergent behaviours in about 37% of the tests. Test regime 1 only sees the effect of a limited set of different settings and does not trigger as much emergent behaviours as either of the other two alternatives. We consider Test regime 1 not satisfactory for detection of the emergent behaviours in question. Since it is only capable of detecting emergent behaviours in about 3% of the tests.

If we are to select only a few "rainy day" scenarios to complement "sunny day" verification testing, we should choose test cases with factor C at high level to ensure the test regime will detect the emergent behaviour type F2. This is the least probable behaviour to detect, based on the effect of factors found in Section 2.3.1. We should further include some test cases with factor B at high level, factor A at low level, and factor D at low level.

In all statistical inference, we use an idealized model to approximate a real-world process that interests us (Lambert, 2018). The model for exploring probabilities in this paper is no exception, leaving some residual risk for the operational phase of the product.

5 Conclusion

The results show that the company could benefit from changing to an alternative test regime, which has higher probability of detecting a given set of unwanted behaviours emerging through system integration testing. The current test regime does not sufficiently trigger the emergent behaviours explored in this paper, but an alternative test regime indicates that the company should be able to sufficiently detect the given set of emergent behaviours.

6 Further Work

The company must perform further analysis to find the optimum test regime to meet all the requirements considering the different needs from integration, verification, and validation testing.

Acknowledgements

We would like to thank Kent Aleksander Kjeldaas for data input to make the paper relevant to company specific problems.

Notations

Table 6. Nomenclature.

Notation	Description
[0…1]	Not present (0) or present (1)
A	Factor A at high level [+]
A'	Factor A at low level [-]
β_A	Coefficient of factor A
F1	Emergent behaviour type 1
G	Any emergent behaviour type
P(A)	Probability of factor A at high level
P(A')	Probability of factor A at low level
P(AB)	Probability of both factor A and B at high level
P(F1)	Probability of emergent behaviour type 1
T1	Test regime 1

References

R. Allenby and A. Slomson. *How to count: An introduction to combinatorics*. CRC Press. 2010.

C. Cenedese and A. L. Gordon. *Ocean Current*. Encyclopedia Britannica. 2021. https://www.britannica.com/science/ocean-current.

Concept Draw. 2021. https://www.conceptdraw.com/a2059c3/p3/preview/640/pict--4-set-venn-diagram-vector-stencils-library.

K. Dunn. Process Improvements Using Data. 2021. https://learnche.org/pid.

R. A. Haugen and M. Mansouri. Applying Systems Thinking to Frame and Explore a Test System for Product Verification; a Case Study in Large Defence Projects. In *Proceedings – 30th INCOSE International Symposium, INCOSE IS 2020, 20-22 July, 2020, virtual*, pages 78-93, 2020. doi:10.1002/j.2334-5837.2020.00709.x

K. A. Kjeldaas, R. A. Haugen, and E. Syverud. Challenges in Detecting Emergent Behavior in System Testing. In *Proceedings – 31st INCOSE International Symposium, INCOSE IS 2021, 17-22 July, 2021, virtual*, pages 1211-1228, 2021. doi: 10.1002/j.2334-5837.2021.00896.x

B. Lambert. *A Student's Guide to Bayesian Statistics, 1st ed*. SAGE Publications Ltd. 2018.

S. Mittal, S. Diallo, and A. Tolk. *Emergent Behavior in Complex Systems Engineering: A Modelling and Simulation Approach*. Wiley. 2018.

D. C. Montgomery. *Design and Analysis of Experiments, 8th ed*. Wiley. 2017.

D. S. Sivia and J. Skilling. *Data Analysis: A Bayesian tutorial, 2nd ed*. Oxford Science Publications. 2006.