Enhanced Data Driven Decision Support

Chris Rijsdijk[1], Tiedo Tinga[2]

[1,2]*Netherlands Defence Academy, Military Technical Sciences, Den Helder, The Netherlands*
c.rijsdijk.01@mindef.nl
t.tinga@mindef.nl

[2]*University of Twente, Dynamics based Maintenance group, Enschede, The Netherlands*
t.tinga@utwente.nl

ABSTRACT

As items are increasingly being equipped with sensors, the applicability of data driven decision support will similarly grow. This paper surveys an endeavor to support decisions with sensor recordings that were coincidentally available. To become meaningful decision support, these sensor recordings should enable better causal inferences because decisions are intended to cause the future. However, data driven decision support is not trivial as normative decision theory is known to suffer from validation issues. This work attempts to alleviate concerns about (*i*) the assessment of preference, (*ii*) causal inferences from non-experimental data and (*iii*) the assessment of the uncertainty about the prospective outcome of a decision. This work will demonstrate that sensor recordings indeed can provide appreciable decision support by presenting two typical cases of human recorded events that were enriched with sensor recordings. From these sensor recordings, prima facie causes and effects of a decision maker's concern were inferred. These type of inferences may potentially have a considerable impact on conventional maintenance policy assessments following a reliability centered maintenance process. Reliability centered maintenance merely anticipates on the *believed consequences of failures* by scheduled inspections, overhauls or discards. As sensor recordings are efficiently collected at a high sampling rate, scheduling inspections may become superfluous. Sight on the prima facie *causes of failures* may enable a kind of proactive control of failures that has not been addressed in the decision logic of a reliability centered maintenance process.

1. INTRODUCTION

As sensor recordings become common practice, their impact on decision making could be huge if present. Therefore, this

work specifically concentrates on the potential contribution of sensor recordings to decision support. Evidently, sensor recordings by themselves will not resolve the validation issues of normative decision theory, but they may alleviate some concerns. This work will explore to what extent sensor recordings may improve (*i*) preference assessments, (*ii*) causal inferences and (*iii*) probability estimations in some typical case studies.

This work will only use sensor recordings that are coincidentally available. So, this work takes an operational rather than a design perspective.

This work is organized as follows: Section 2 will introduce three generic concerns about data driven decision support. Section 3 will outline why sensor recordings may alleviate these concerns. Section 4 will present a realistic case of data driven decision support. Finally, Section 5 will discuss the findings and present some conclusions.

2. BACKGROUND

This Section will introduce three generic concerns about data driven decision support.

2.1. Preference assessments

Decision makers are typically not indifferent towards their choices, i.e. they typically have some preference. So, any decision to act is somehow preferred over the decision not to act. If preference would exclusively reside in an individual's mind, decision making would lose much of its importance to society. However, concerns about ways to substantiate preference with some observable utility attributes have a long history as illustrated by the St. Petersburg paradox (Bernoulli, 1954 [1738]).

In case of group decisions, common sense may alleviate concerns about these utility attributes. As an organization can only exist by the choice to collaborate, individuals should align their individual utilities. To enable this alignment, the group's utility should somehow become

1

explicit to these individuals. This work posits that lagging performance indicators (Drucker, 1954), (Kaplan & Norton, 1996) used within an organization represent that organization's common sense about utility attributes. Note that conventional lagging performance indicators, like for example reliability, do typically not allow for causal inferences that may support decisions to control them. Respecting some construction rules for performance indicators (Rijsdijk & Tinga, 2016) contributes to a decision maker's ability to infer their causes from recording routines. To illustrate this point, Section 4 will show that an organization's conventional performance indicators may be improved.

2.2. Causal inferences

Associated symptoms may already suffice for predictions, but causes could also point at some means to control. If these causes could be inferred from (non-experimental) recording routines, i.e. a causal relation between a lagging performance indicator and a specific (sensor) recording could be established, a decision maker could also learn *how* to control the future. Otherwise, he may just anticipate by expert judgement.

This work adopts Granger's (1980) proposition that a cause C_T (at time T) entails unique information about the effect E_{T+1} (at time $T+1$) that is not available otherwise. So eliminating C_T from the set of all information in the universe up to now (Ω_T) matters for E_{T+1}:

$$Pr(E_{T+1}|\Omega_T) \neq Pr(E_{T+1}|\Omega_T/C_T) \qquad (1)$$

Equation (1) implements the principle that an effect cannot precede a cause in time which prohibits E_{T+1} to cause any element from Ω_T. Random assignment of C_T treatments could help to assess Eq. (1), because C_T would have been the only variable that could eventually associate with E_{T+1}. In practice, however, a decision maker does not have the possibility to test all kinds of conditions, but has access to only some non-experimentally collected subset of Ω_T. Then, the decision maker simply cannot infer the causality $C_T \rightarrow E_{T+1}$ in Eq. (1) from only the available recording routines. Still, the decision maker may already appreciate a modest notion of prima facie (~ *at first sight*) causality that only holds with respect to some finite information set *V*. For example C_T prima facie causes E_{T+1} with respect to the information set $V=\{e_{t+1},c_t\}$ if:

$$Pr(E_{T+1}|C_T) \neq Pr(E_{T+1}) \qquad (2)$$

In a limited universe where each $V=\{e_{t+1},c_t\}$ is seen as a replication, the decision maker may infer the likelihood of Eq. (2) from recording routines. However, an extension of the information set to $V=\{e_{t+1},c_t,b\}$ may reduce the prima facie causality in Eq. (2) to a spurious causality:

$$Pr(E_{T+1}|C_T, B) = Pr(E_{T+1}|B) \qquad (3)$$

In Eq. (3), C_T appears in fact independent of E_{T+1} despite the prima facie causality in Eq. (2) because *B* could act as a confounder $B \rightarrow (E_{T+1}, C_T)$ or as a mediator $C_T \rightarrow B \rightarrow E_{T+1}$. If *B* is a confounder, control over *B* affects E_{T+1} and control over C_T would not affect E_{T+1}. If *B* is a mediator, control over both *B* and C_T affects E_{T+1} but control over *B* would pre-empt control over C_T. Still, a decision maker who wants to get insight in some *prospective* effect E_{T+1} while being unaware of *B*, may already appreciate knowledge about the prima facie causality in Eq.(2).

Now it is clear that inferring (prima facie) causality is useful for decision making, the next problem is that the inference of causalities from recording routines in practice is problematic. This is the case for two reasons: (*i*) the effect of background variables cannot be managed by random assignment of treatments and (*ii*) the decision maker has no control over the composition of the sample. This sample could therefore just fail to satisfy the following preconditions to infer the modest notion of prima facie causality in Eq. (2). *Firstly*, E_{T+1} and C_T must be sampled at a rate that allows to reconstruct their original signal and *secondly*, C_T must vary because the effect of a constant cause remains unobserved (ceteris paribus). In some cases, data driven decision support may appear to be inapplicable because the recording routines just fail to satisfy these preconditions.

The independence assumptions required to interpret a prima facie causality as causal may straightforwardly be specified by Pearl's (2010) structural causal model. These independence assumptions may appear to be highly problematic, but the structural causal model is explicit about them at least.
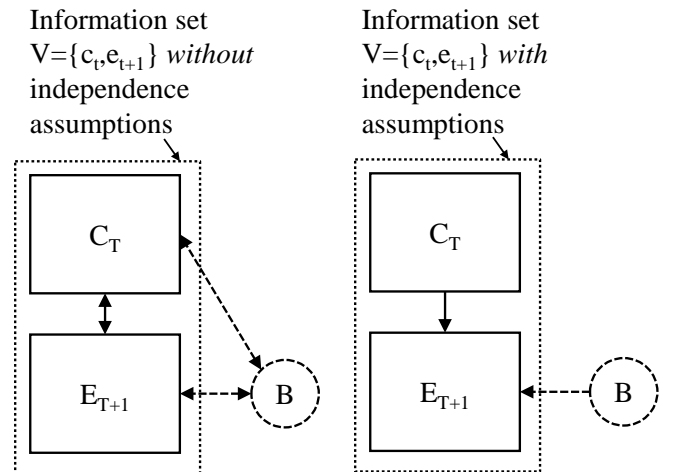

Figure 1. Examples of structural causal models

The independence assumptions to interpret the prima facie cause in Eq. (2) as causal are represented by the missing arrows between *B*, C_T and E_{T+1} in the structural causal model of Fig. 1:

$E_{T+1} \nrightarrow C_T$ This independence assumption is not problematic as the future cannot cause the past by common sense (Granger, 1980).

$C_T \nleftrightarrow B$ These independence assumptions imply the non-existence of mediators $C_T \rightarrow B \rightarrow E_{T+1}$ or confounders $B \rightarrow (C_T, E_{T+1})$ which is highly problematic as B could be any element of the universe Ω_{T+1}.

Even if a prima facie causality appears to be inferable at all, it merely remains inconclusive decision support. Section 4 will illustrate the inference of a prima facie causality from sensor recordings that inconclusively directs a decision maker to some means to control a prospective effect.

2.3. Probability assessments

The third and final concern about data driven decision support is the assessment of the probability of the various prospects a decision maker has, which are typically uncertain. This uncertainty is predominantly quantified probabilistically. The axioms of probability theory (Kolmogorov, 1933) are widely accepted. However, the interpretation of a probability appears to be controversial. Even subjectivists who consider a probability as some degree of rational belief, still tend to use past frequencies to tie a probability to an observable reality. Past frequencies are assignable to the various outcomes of a replication. Replications are random experiments which are *presumed* to be identical. A random experiment could be anything that yields an uncertain outcome of interest.

This work posits that a decision maker can only influence the yet-to-be-observed future but at the moment of deciding, he can only *believe* in some future. Observed frequencies on the other hand are only retrievable from the past. Observed frequencies will therefore never be entirely compelling for the future. A probabilistic approach to data driven decision support therefore requires both a subjective and a frequency interpretation of a probability. To be more specific, a probabilistic approach to data driven decision support requires (*i*) an *arbitrary* criterion to identify sufficient replications among the recordings and (*ii*) an *arbitrary* belief that the future will also be a replication. The controversy about these presumptions has typically been left unquantified. Moreover, data cannot support decisions that are *believed* to be unprecedented.

3. ENHANCED DATA DRIVEN DECISION SUPPORT

This Section will outline why sensor recordings may enhance data driven decision support by alleviating the three concerns from Section 2. To illustrate the point, two scenarios of data driven decision support will be compared. Scenario 1 entails decision support from human recorded events (actions) that rely on expert judgement. Scenario 2

augments scenario 1 with sensor recordings representing a concern of a decision maker.

A decision maker may appreciate scenario 2 because sensor recordings typically better comply with the three construction rules for performance indicators (Rijsdijk & Tinga, 2016): indicators should be independent and non-redundant, should be sampled at sufficiently high rates and should entail a sufficient number of replications, as will be discussed next.

Firstly, sensor recordings avoid redundancy since they reflect a state of affairs at a discrete point of time. A dependency between adjacent sensor recordings therefore has a *causal* and not a *definitional* explanation. Conventional performance indicators that have been built on human recorded events are often times-to-event (e.g. time to failure) or event rates (e.g. failure rate) that are only observable over a time interval. If this time interval exceeds the sampling interval, consecutive conventional performance indicators would become dependent *by definition* because they would partially rely on the same events.

Secondly, sensor recordings are efficiently sampled at a high rate that may better reconstruct reality. Conventional performance indicators are built on recordings of expert judgement that does not explicitly relate to reality.

Finally, as sensor recordings are non-redundant and sampled at a higher rate, causal models may be inferable *faster*. These causal models (and *not* definitional dependencies!) are the essential data driven support that may strengthen a belief in some unobserved prospect.
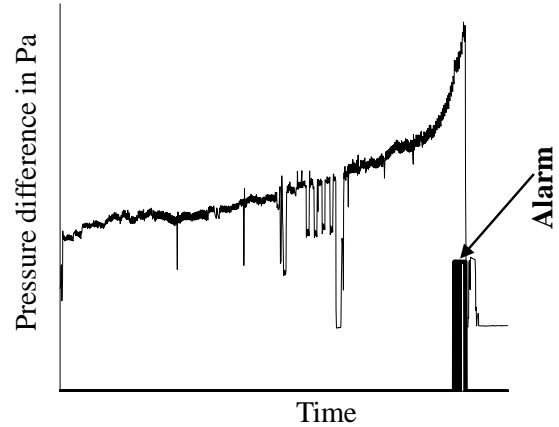
Figure 2. Evolution of clogging

Figure 2 depicts the evolution of some pressure difference as a fouling indicator that detects a flow restriction. Once this fouling indicator exceeds 20 Pa, a decision to clean will be triggered. In scenario 1, the decision maker only has

access to the human recorded cleaning events in a computerized maintenance management system. In scenario 2, the decision maker has *also* access to the sensor recordings of the fouling indicator. The remainder of this Section will compare these two scenarios with respect to the three concerns raised in subsections 2.1 to 2.3.

3.1. Preference assessments

Let the cleaning effort C (in this case the cause) and the unrestricted flow E (the effect) be the utility attributes that substantiate the decision maker's preference to clean or not to clean at particular times.

In scenario 1, the decision maker only has recordings of the cleaning events C. Then, just one of his utility attributes has been recorded. So, the decision maker can only *believe* in some flow restriction E.

In scenario 2, the decision maker has *also* access to sensor recordings of an indicator for the flow restriction E. Then, both of his utility attributes have been recorded. So, the decision maker may produce a plot like Fig. 2 that depicts both the cleaning events C and the indicator for the flow restriction E. This allows the decision maker to more precisely substantiate his utility of the past.

In conclusion, sensor recordings may enhance preference assessments by providing a more complete set of utility attributes that rely on physical measurements rather than on subjective expert judgement.

3.2. Causal inferences

Let the decision maker *believe* that a cleaning event C_T *causes* an unrestricted flow E_{T+1}.

In scenario 1, the decision maker has only access to recordings of cleaning events. Therefore, the decision maker's belief in the causality $C_T \rightarrow E_{T+1}$ *cannot* be substantiated by recording routines and the flow effect E_{T+1} of the cleaning events C_T remains unrecorded if observed at all.

Still, the recordings of the past cleaning events C_T may enable better predictions of the next cleaning event. However, time by itself is expected to be merely an uncontrollable associated variable rather than a cause of cleaning. Therefore, scenario 1 may enable a decision maker to better *anticipate* on the next cleaning event (e.g. by scheduling resources) but the causes or effects of cleaning remain unrecorded. Then, the use of cleaning remains a belief. Nevertheless, time is important to a decision maker because an allowance to defer a cleaning event till infinity is practically equivalent to no cleaning at all. Eventually, time may even *define* cleaning events if the decision maker adopts a time based maintenance policy. So, although time does generally not cause failures, time remains an essential element of any policy.

In scenario 2, the decision maker has access to the recordings of both the cleaning events C_T and the fouling indicator E_{T+1} as depicted in Fig. 2. As opposed to scenario 1, the decision maker's belief in the causality $C_T \rightarrow E_{T+1}$ can be substantiated by recording routines. So, the prima facie effect E of a cleaning event C in Eq. (2) may be inferable. As the decision maker accepts the independence assumptions in Fig. 1, he may control the prospective fouling indicator E_{T+1} by a (properly timed) cleaning event C_T using the prima facie causality in Eq. (2) that has been inferred from recording routines.

In scenario 2, the decision maker may also verify cleaning policy compliance. The cleaning policy *defines* that the fouling indicator E triggers a cleaning event C when it surpasses some limit. As opposed to scenario 1, a posterior verification of the cleaning policy compliance is possible in scenario 2.

As the fouling indicator E tends to steadily drift towards the alarm limit, the current fouling indicator E may appear to be a much better predictor of the next cleaning event C than the past cleaning events C. In scenario 2, a decision maker has more options to better *predict* the next cleaning event C.

The fouling indicator E in Fig. 2 reveals some spikes and discontinuities that evoke a quest for a cause but these causes remain unrecorded and would require expert judgement, even in scenario 2.

In conclusion, inferences of prima facie causes from sensor recordings may strengthen a belief in an observable effect of decisions. Moreover, sensor recordings may verify policy compliance and enable prediction of events.

3.3. Probability assessments

The 3rd and final concern was the assessment of probabilities of a decision makers prospects. Let the decision maker's uncertainty about the future be quantified probabilistically. To assign a probability to a prospect from past frequencies, a decision maker should arbitrarily presume (*i*) a criterion to identify sufficient replications and (*ii*) that the prospect will also be a replication. Too often, replications remain unidentifiable because the recording routines are incomplete.

In scenario 1, the decision maker only has recordings of the cleaning events C. This delimits the observable criteria to identify replications among these cleaning events to their time stamp. Eventually, cleaning events associate with time (e.g. Duane plots or Fig. 3) or with life time (e.g. Weibull plots). As (life) times are not instantaneously observable, collecting sufficient replications typically requires much time (Abernethy, 2006) and the spatiotemporal proximity of a causality may remain unobserved (Rijsdijk, 2016). As time is merely seen as an uncontrollable associated variable rather than as a cause of cleaning events, inferred (life) time

dependencies only enable a decision maker to *predict* cleaning events.

Alternatively, the decision maker may adopt some hazard rate model while assuming that the number of cleaning events at each time interval is a replication. Then, replications may be collected more efficiently but the cleaning events are presumed to be independent of (life) times. Section 4.2 will use a hazard rate model to test whether events arrive independent of time. Reliability data handbooks similarly presume time independence (NSWC, 2011), (DoD, 1991) provided that given conditions have been satisfied. Given conditions are an important component in any reliability definition (IEC, 1990), but in scenario 1, satisfaction of these given conditions cannot be established by recordings. Therefore, the hazard rates in reliability data handbooks (NSWC, 2011), (DoD, 1991) cannot be evaluated in scenario 1.

In scenario 2, the decision maker may better approximate his desired criterion to identify replications. For example, the decision maker may censor cleaning events from being a replication because they were not triggered by the surpassing of the alarm limit of the fouling indicator. These 'rejected' cleaning events may be seen as cases of opportunistic maintenance that were triggered by other failures than excessive fouling. In scenario 1, this censoring would have been impossible.

In scenario 2, the decision maker has also access to sensor recordings of the fouling indicator. As compared to human entered recordings, sensor recordings more efficiently yield replications as they (*i*) omit the human effort of recording and they (*ii*) are typically sampled at a higher rate. Therefore, the sensor recordings in Fig. 2 may provide a faster feedback to a guessing decision maker. A decision maker may for example guess that some redesign would cause the fouling to reduce. The evolution of the fouling indicator during the first time-to-clean after this redesign may well suffice to test this guess *rapidly*. In scenario 1, the decision maker would have been required to await for several times-to-clean to test his guess. Sensor recordings may therefore make responses to decisions faster observable to impatient decision makers.

In conclusion, sensor recordings can enhance probability assessments because they are more efficiently collected, i.e. they yield more candidate replications in a shorter time. Moreover, sensor recordings can identify replications by observations rather than by (recorded) expert judgement.

4. ENHANCING DATA DRIVEN DECISION SUPPORT

This Section will portray a fleet operator who developed his data driven decision support from scenario 1 to scenario 2 as introduced in Section 3. Section 4.1 will outline the fleet operator's conventional performance indicators that have been built on human recorded events that rely on expert judgement (scenario 1). Section 4.2 will apply the construction rules from Rijsdijk and Tinga (2016) to scenario 1. Section 4.3 will infer a prima facie cause of a lagging performance indicator that is only possible in scenario 2. Eventually, the expected benefits of sensor recordings to decision support reveal.

4.1. Conventional data driven decision support

The performance indicators of this fleet operator resemble the common sense of Blanchard (2004) and Jones (2007). To illustrate the point, this work confines to a reliability indicator R:

$$R = \frac{1}{MTBF} = \frac{cumulative\ failures}{cumulative\ deployment\ time} \tag{4}$$

where the mean time between failures *MTBF* followed from the cumulative deployment time of the fleet divided by the cumulative failures. As time evolves and the denominator of Eq. (4) grows, this indicator R gradually behaves more stationary and the next step ahead value of the indicator R tends to become better predictable. However, this predictability relies on the amount of past deployment time that can no longer be controlled. For a decision maker who pursues reliability, only the future is potentially controllable. However, the indicator R is just a measure of tendency that has been built on the past. Therefore, the indicator R poorly supports predictions that are meaningful to a decision maker, but it may well enable an assessment of posterior compliance with some requirement, while levelling out outliers over time. The fleet operator actually used its performance indicators to identify nonconformities that were addressed by expert judgement. The fleet operator did not apply causal inferences from recording routines to predict, let alone control performance indicators.

4.2. Decision support from human recorded events

Rijsdijk and Tinga (2015) proposed to replace a redundant indicator R by an instantaneously observable variable that addresses the same concern, as is shown in Figure 3. Evidently, Fig. 3 just presents the numerator as a function of the denominator of Eq. (3), but it much better shows the fluctuations in the cumulative failures over time. Fluctuations rather than steadiness allow a decision maker to learn about the system behavior. Figure 3 is just an alternative representation of scenario 1 that respects the construction rules for performance indicators (Rijsdijk & Tinga, 2016). In scenario 1, the fleet operator ignores sensor recordings as a criterion to identify replications. A decision maker could then only resort to some principle of insufficient reason by *defining* each discrete time interval in Fig. 3 as a replication.
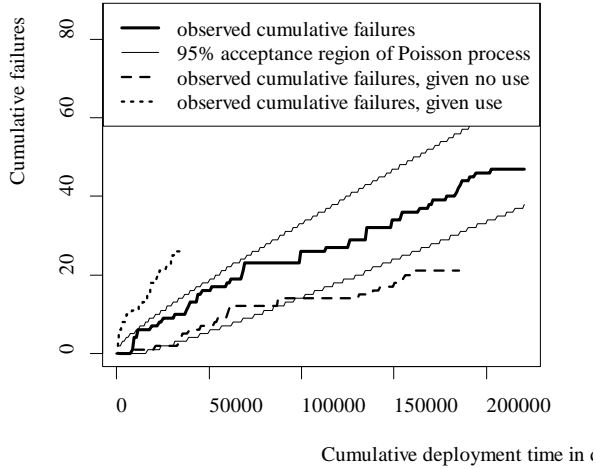
Figure 3. Evolution of the cumulative failures

The fleet operator may for example assume that the number of failures k at every discrete time interval is a replication from a Poisson distribution with parameter λ:

$$Pr(K = k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{5}$$

Then, the expectation and the variance of the number of failures $E[K]$ at every discrete time interval follows from:

$$E[K] = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda$$

$$Var(K) = E[K^2] - (E[K])^2 = \lambda \tag{6}$$

Equation (6) only holds for a single replication K, but the sum of t replications follows from:

$$E\left[\sum_{i=1}^{t} K_i\right] \xrightarrow{i.i.d.} t \times E[K] = t \times \lambda$$

$$Var\left(\sum_{i=1}^{t} K_i\right) \xrightarrow{i.i.d.} t \times Var(K) = t \times \lambda \tag{7}$$

Equation (7) shows that the expectation and the variance of the cumulative failures during t discrete time intervals linearly grow in time. Moreover, the cumulative failures over t intervals may again be seen as a replication from a Poisson distribution with a parameter $t\lambda$. Figure 3 depicts the 95% acceptance region of these Poisson distributions where λ has arbitrarily been estimated by ordinary least squares regression:

$$\frac{\delta \sum_{i=1}^{t}(k_i - i \times \lambda)^2}{\delta \lambda} = 0 \rightarrow \lambda = \frac{\sum_{i=1}^{t} k_i \times i}{\sum_{i=1}^{t} i^2} \tag{8}$$

An acceptance region defines the upper and the lower bounds of the observations, given some presumed definition

of a replication. As the observed cumulative failures in Fig. 3 evolve within the acceptance region, the presumed definition of a replication has not been rejected. Further analyses revealed that over 80% of the failure modes at this fleet operator similarly arrived within a 95% acceptance region. This result, confirming that failures generally arrive independent of time, may not be very surprising since time by itself is unlikely to cause a failure mode. The few failure modes that did associate with time typically require a quest for a mediating or a confounding cause. Although rarely refuted on statistical grounds, the presumption that failures arrive randomly in time is problematic for control over failures. Resembling a fair casino, prospective failures then become a matter of destiny that cannot be controlled. The intuition here is that failures have causes that may be controllable. Eventually, the fleet operator may infer these causes from recording routines. Potentially, the fleet operator may benefit from these inferred prima facie causalities, even in case a presumption of randomly arriving failures has not been rejected. Still, better predictions of inevitable failures may already enable the fleet operator to better anticipate on their *believed consequences* as advocated by Nowlan and Heap (1978) and Moubray (1997).

This section just applied the construction rules of Rijsdijk and Tinga (2016) to performance indicators that have been built on human recorded expert judgement (scenario 1). As opposed to Section 4.1, the fleet operator arbitrarily defined a replication which allowed him to only *predict* failures that seem to arrive randomly.

4.3. Decision support from human recorded events and sensor recordings

In this case, the fleet operator *believed* that deployment caused the failures in Fig. 3. As the gearbox time of each *individual* fleet member has been recorded, a deployment indicator becomes accessible. Let this deployment indicator be a dichotomous variable C_T:

$$C_T = \begin{cases} 0; & no\ gearbox\ time\ between\ t\ and\ t+1 \\ 1; & gearbox\ time\ between\ t\ and\ t+1 \end{cases} \tag{9}$$

Fortunately, the failures in Fig. 3 were also traceable to *individual* systems (fleet members), which contributes to the spatiotemporal proximity of a cause and its effect, i.e. deployment and failures could be related *within* fleet members and not just *across* the fleet. The ability to test for a prima facie causality within fleet members also increases the sample size. The cumulative deployment time in Fig. 3 just entails a sample of 210 fleet days which is equivalent to a sample of 220500 fleet member days because there are 1050 fleet members.

Let the functionality of an individual fleet member be a dichotomous variable E_{T+1}:

6

$$E_{T+1} = \begin{cases} 0; \; no \; failure \; between \; t+1 \; and \; t+2 \\ 1; \; failure(s) \; between \; t+1 \; and \; t+2 \end{cases} \quad (10)$$

Let the sampling rate of the variables C_T and E_{T+1} be daily. As E_{T+1} is only human recorded expert judgement, some randomness in the delay of recording prevents a higher sampling rate to reconstruct the signal better. If a failure could have been *defined* in terms of a sensor recording similar to Fig. 2, a higher sampling rate might have allowed for a better signal reconstruction.

Let C_T *not* prima facie cause E_{T+1} with respect to the information set $V=\{c_t, e_{t+1}\}$. Then, each $\{e_{t+1}\}$ has been *defined* as a replication from a Bernoulli distribution and the likelihood of observing a number of $(m_0 + m_1)$ failures in the sample of $(n_0 + n_1 = 220500)$ information sets V follows from:

$$pr\big(m_0 + m_1 | n_0 + n_1, Bern(p)\big) = \binom{n_0 + n_1}{m_0 + m_1} p^{m_0 + m_1} (1-p)^{n_0 + n_1 - m_0 - m_1} \quad (11)$$

Moreover, the likelihood of m_0 observed failures among n_0 replications where $C_T = 0$ and of m_1 observed failures among n_1 replications where $C_T = 1$ is then given by:

$$pr\big(m_0 | n_0, Bern(p)\big) \times pr\big(m_1 | n_1, Bern(p)\big) = \binom{n_0}{m_0} p^{m_0} (1-p)^{n_0 - m_0} \binom{n_1}{m_1} p^{m_1} (1-p)^{n_1 - m_1} \quad (12)$$

Given Eq. (11), the likelihood of Eq. (12) follows from Fisher's exact conditional approach:

$$pr(m_0, m_1 | Bern(p), n_0, n_1, m_0 + m_1) =$$
$$\frac{Eq.(12)}{Eq.(11)} = \frac{\binom{n_0}{m_0} \times \binom{n_1}{m_1}}{\binom{n_0 + n_1}{m_0 + m_1}} \quad (13)$$

The likelihood in Eq. (13) directly follows from the sample as the Bernoulli parameters in Eq. (12) and Eq. (11) cancel out. The statistical significance of Eq. (13) follows from the probability value. The probability value is the probability of obtaining an outcome that is at least as extreme as the outcome being observed. Rijsdijk (2016) proofed, without using some Wald statistic, that this probability value directly follows from:

$$p \; value =$$
$$\begin{cases} \sum_{\forall c \ge m_0} \frac{\binom{n_0}{c}\binom{n_1}{m_0 + m_1 - c}}{\binom{n_0 + n_1}{m_0 + m_1}} \; if \left(\frac{m_0}{n_0}\right) \ge \left(\frac{m_0 + m_1 - 1}{n_0 + n_1}\right) \\ \sum_{\forall c \ge m_1} \frac{\binom{n_1}{c}\binom{n_0}{m_0 + m_1 - c}}{\binom{n_0 + n_1}{m_0 + m_1}} \; if \left(\frac{m_1}{n_1}\right) \ge \left(\frac{m_0 + m_1 - 1}{n_0 + n_1}\right) \end{cases} \quad (14)$$

This implies that, if the probability value is *above* an arbitrarily chosen significance level α, observing more extreme frequencies is likely and the presumed non prima facie causality that underlies Eq. (13) will *not* be rejected. If the probability value is *below* the significance level α, observing more extreme frequencies is unlikely and the presumed non prima facie causality that underlies Eq. (13) will be rejected.

		C_T	
		0 (not deployed)	1 (deployed)
E_{T+1}	0 (no fail)	185236	35217
	1 (fail)	21	26

Table 1. Contingency table

For the observed frequencies in this case study, as shown in Table 1, the probability value follows from:

$$p \; value = \sum_{\forall c \ge 26} \frac{\binom{35243}{c} \times \binom{185257}{47 - c}}{\binom{220500}{47}} \approx 7 \times 10^{-10} \quad (15)$$

Equation (15) immediately follows from the observed frequencies in Table 1 without any estimation of a Bernoulli parameter. The probability value in Eq. (15) is far below a typical significance level of $\alpha = 0,05$ which rejects the presumed *non* (prima facie) causality that underlies Eq. (13). This rejection only supports the initial *belief* of the fleet operator that C_T causes E_{T+1}, as the additional independence assumptions from Fig. 1 are required to interpret this prima facie causality as a causality $C_T \rightarrow E_{T+1}$. The fleet operator may decide to accept these independence assumptions and he then can decide to control the functionality E_{T+1} by the deployment indicator C_T. Otherwise, the fleet operator could decide to at least better predict the functionality E_{T+1} by knowledge about the deployment indicator C_T.

The two dashed lines in Figure 3 confirm Eq. (15) by showing that a partitioning of the cumulative deployment time by the deployment indicator C_T yields observed cumulative failures that are outside the 95% acceptance region of the Poisson distributions with parameters $t\lambda$. This means that the estimated λ appeared to be acceptable in scenario 1, but not in scenario 2. The fleet operator can now consider to estimate specific Poisson parameters λ for the cumulative failures given deployment $C_T = 1$ and given no deployment $C_T = 0$ by ordinary least squares regression Eq. (8).

In summary, this Section inferred a prima facie causality that enabled a better functionality E_{T+1} prediction from a deployment indicator C_T. In scenario 2, the fleet operator could "prima facie" control prospective functionality E_{T+1} by the deployment indicator C_T. The fleet operator would not have been able to do this in scenario 1 while using his conventional performance indicators. Still, the functionality E_{T+1} remained recorded subjective expert judgement. A sensor recording indicating functionality would have tied this case better to an observable reality.

5. DISCUSSION AND CONCLUSION

This work attempted to enhance data driven decision support by alleviating concerns about (*i*) the assessment of preference, about (*ii*) causal inferences from non-experimental data and about (*iii*) the assessment of the uncertainty about the prospective outcome of a decision. Two typical examples of sensor recordings have been used to demonstrate that decision support can indeed be enhanced by (*i*) making the common sense about (an organization's) preference observable, by (*ii*) making the effects of decisions faster observable and by (*iii*) generating more candidate replications in a shorter time and at a lower effort.

The first case (Figure 2) showed that sensor recordings also allowed to verify policy compliance, as the cleaning events were known to be triggered by the fouling indicator (condition based maintenance). The failures in the second case (Figure 3) remained human recorded expert judgement, that could not be verified a posteriori. Rijsdijk (2018) recently reviewed some common sense approaches to observe functionality from sensor readings, but also concluded that functionality assessments are typically problematic.

Causal inferences as illustrated in this work may have a huge impact on conventional maintenance policy assessments by a typical reliability centered maintenance process. Reliability centered maintenance merely anticipates on the *believed consequences of failures* by scheduled inspections, overhauls or discards. As sensor recordings are efficiently collected at a high sampling rate, scheduling inspections may become superfluous. Sight on the prima facie *causes of failures* may enable a kind of proactive control of failures that has not been addressed in the decision logic of a reliability centered maintenance process.

This work not only confirmed that the construction rules for performance indicators (Rijsdijk & Tinga, 2016) enable predictions of interest to a decision maker (Section 4.1 compared with Section 4.2) but also that they actually enable the inference of a prima facie cause (Section 4.3). This prima facie cause followed from Fisher's exact conditional approach Eq. (15). The motivation for Fisher's exact conditional approach has been detailed in (Rijsdijk, 2016).

Data driven decision support should not be seen as some panacea. *Firstly*, the utility attributes that substantiate a decision maker's preference may lack common sense or they may remain unrecorded. So, the generic concerns from Section 2.1 remain unresolved, but sensor recordings can certainly provide a more complete set of observable utility attributes. *Secondly*, causal inferences from recording routines remain problematic because the effect of background variables cannot be managed by random assignment of treatments and the decision maker has no control over the composition of the sample. This means that the requirements on the composition of the sample may simply not be met. For example, the effect of a constant cause cannot be inferred from data (ceteris paribus). So, the generic concerns from Section 2.2 also remain unresolved, but again sensor recordings can surely strengthen a belief in an observable effect of decisions. *Finally*, a probabilistic quantification of the uncertainty in a specific prospect may fail because (*i*) the criterion to identify sufficient replications appears to be controversial or because (*ii*) the prospect is believed to be unprecedented. Again, these generic concerns from Section 2.3 remain unresolved, but sensor recordings do have the potential to enhance probability assessments because they are more efficiently collected.

In conclusion, the two cases in this work have shown that sensor recordings can provide observable explanations for an organization's performance indicators that are conventionally built on human recorded events. Decision makers may then test their beliefs about how to control these performance indicators at some degree of certainty.

REFERENCES

Abernethy, R.B. (2006). *The New Weibull Handbook; reliability and statistical analysis for predicting life, safety, survivability, risk, cost and warranty claims.* North Palm Beach: Abernethy.

Bernoulli, D. (1954 [1738]). Exposition of a new theory on the measurement of risk. *Econometrica,* vol.22, pp.23-36.

Blanchard, B.S. (2004). *Logistics engineering and management.* Upper Saddle River: Prentice Hall.

Department of Defense (DoD) (1991). Reliability Prediction of Electronic Equipment. In MIL-HDBK, *MIL-HDBK-217F,* Washington: Department of Defense.

Drucker, P. (1954). *The practice of management.* New York: Harper & Row.

Granger, C. (1980). Testing for causality, a personal viewpoint. *Journal of Economic Dynamics and Control,* Vol.2 Nr.1, pp.329-352.

International Electrotechnical Committee (IEC) (1990). International Electrotechnical Vocabulary, chapter 191: Dependability and quality of service. In IEC, *IEC 60050-191;* Geneva: International Electrotechnical Commission.

Jones, J.V. (2007). *Supportability engineering handbook: implementation measurement and management.* New York: Sole Press.

Kaplan, R., & Norton, D. (1996). *The balanced scorecard; translating strategy into action.* Boston: Harvard Business School Press.

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Berlin: Springer.

Moubray, J. (1997). *Reliability Centred Maintenance.* New York: Industrial Press.

Nowlan, F., & Heap, H. (1978). *Reliability Centered Maintenance.* Los Altos: Dolby Access Press.

NSWC. (2011). *Handbook of Reliability Prediction Procedures for Mechanical Equipment.* West Bethesda: Naval Surface Warfare Center.

Pearl, J. (2010). An introduction to causal inference. *International Journal of Biostatistics,* Vol.6, Nr.2. doi:10.2202/1557-4679.1203.

Rijsdijk, C. (2016). *Maintenance is unjustifiable; an improved inference.* Doctoral dissertation. Twente University, Enschede.

Rijsdijk, C. (2018). *Observing functionality; a decision maker's concern.* Draft in press.

Rijsdijk, C., & Tinga, T. (2015). Enabling maintenance performance prediction by improving performance indicators. In: *Safety and reliability of complex engineered systems.* pp.1001-1007. Zurich: CRC Press.

Rijsdijk, C., & Tinga, T. (2016). Observing the effect of a policy; a maintenance case. *Journal of Quality in Maintenance Engineering.* Vol. 22, Nr.3. doi:10.1108/JQME-10-2014-0055.