# URREF Reliability versus Credibility in Information Fusion (STANAG 2511)

Erik Blasch
Kathryn B. Laskey
Anne-Laure Jousselme

Valentina Dragos
Paulo C. G. Costa
Jean Dezert

Originally published as Blasch E., Laskey K.B., Jousselme A.-L., Dragos V., Costa P.C.G., Dezert J., *URREF reliability versus credibility in information fusion (STANAG 2511),* Proc. of Fusion 2013 Int. Conference on Information Fusion, Istanbul, Turkey, July 9-12, 2013, and reprinted with permission.

***Abstract*— *For many operational information fusion systems, both reliability and credibility are evaluation criteria for collected information. The Uncertainty Representation and Reasoning Evaluation Framework (URREF) is a comprehensive ontology that represents measures of uncertainty. URREF supports standards such as the NATO Standardization Agreement (STANAG) 2511, which incorporates categories of reliability and credibility. Reliability has traditionally been assessed for physical machines to support failure analysis. Source reliability of a human can also be assessed. Credibility is associated with a machine process or human assessment of collected evidence for information content. Other related constructs for URREF are data relevance and completeness. In this paper, we seek to develop a mathematical relation of weight of evidence using credibility and reliability as criteria for characterizing uncertainty in information fusion systems.*

***Keywords:*** *Reliability, Credibility, URREF, PCR5, STANAG2511*

## I. INTRODUCTION

Information fusion is based on uncertainty reduction; wherein the International Society of Information Fusion (ISIF) *Evaluation of Techniques of Uncertainty Reasoning Working Group* (ETURWG) has had numerous discussions on definitions of uncertainty. One example is the difference between reliability and credibility, which is called out in NATO STANAG 2511 [1]. To summarize these ETURWG discussions, we detail an analysis of credibility and reliability.

Information fusion consumers comprise users and machines of which the man-machine interface requires understanding of how data is collected, correlated, associated, fused, and reported. Simply stating an uncertainty representation of "confidence" is not complete. From URREF discussions [2]:

*reliability* relates to the source, and
*credibility* refers to the content reported.

There are scenarios in which reliability and credibility need to be differentiated. Examples of information fusion application areas include medical, legal, and military domains. A common theme is involvement of humans in aggregating information. In many situations, there is cause for concern about the reliability of the source that may or may not be providing an accurate and complete representation of credible information. In cases where there is a dispute (e.g., legal), the actors each seek their own interests and thus are asked a series of questions by their own and opposing representations to judge the veracity of their statements.

*Weight of Evidence* (WOE) is addressed in various fields (risk analysis, medical domain, police, legal, and information fusion). In addition to credibility and reliability, *Relevance* assesses how a given uncertainty representation is able to capture whether a given input is related to the problem that was the source of the data request. A final metric to consider is *completeness*, which reflects whether the totality of evidence is sufficient to address the question of interest. These criteria relate to high-level information fusion (HLIF) [3] systems that work at levels three and above of the Data Fusion Information Group (DFIG) model. For the URREF, we then seek a mathematical representation the weight of evidence:

$$\text{WOE} = f(\text{Reliability, Credibility, Relevance, Completeness}) \quad (1)$$

where $f$ is an function to be defined with operations on how to combine such as a utility analysis.

Sect. II. provides related research and Sect. III overviews information fusion. Sect. IV discusses the weight of evidence including relevance and completeness. Sect. V describes the modeling of reliability and credibility with Sect. VI providing a simulation over evidence processing. Sect. VII provides discussion and conclusions.

## II. BACKGROUND

There are many examples of ***reliability*** analysis for system components [4]. Typically, a reliability assessment is conducted on system parts to determine the operational life of each component over the entire collection of parts [5]. A reliability analysis can consist of many attributes such as survivability [6], timeliness, confidence, and throughput [7, 8]; however the most notable is time to failure [9]. Reliability is typically modeled as a continuous analysis of a part; however, a discrete analysis can conducted for the number of failures in a given period of time [10]. Real-time analysis requires information fusion between continuous and discrete analysis over new evidence [11], covariance analysis [12, 13], and resource analysis [14] to control sensors.

To assess the performance of sensors (and operators) requires analysis of the physical reliability of components. Data fusion can aid in fault detection [15], predictive diagnostics [16], situation awareness [17], and system performance. A model of

reliability includes time-dependent measures for operational lifetime analysis and controllability which are aspects of a data fusion performance analysis [18]. Use of multiple systems can aid in reducing failures through redundancy or system reconfiguration in response to failed sensors [19] for such applications as robotics [20, 21], risk analysis for situation awareness [22, 23], and cyber threats [24, 25, 26].

Time-dependent measures such as times between failures are appropriate for processes that operate over time to produce a stream of outputs; and failures can render the output stream unreliable. For systems that respond to discrete queries or produce alerts, such as human operators in a fusion center or pattern recognition systems, reliability is assessed through correspondence between outputs and the actual situation. The confusion matrix (CM) is a typical measure [27]. Reliability also relates to the opinions of observers [28].

*Credibility* To analyze credibility of evidence, we can use probabilistic or credibilistic frameworks such as Bayes, Dempster-Shafer, or following proportional conflict redistribution (PCR) principle, etc. [29, 30, 31]. Credibility of a hypothesis can be assessed through its prior probability or belief; and also through conflict: information is more credible when it does not conflict with other information.

To summarize,

- *Reliability* is an attribute of a sensor or other information source, and measures the consistency of a measure of some phenomenon. Reliability can be assessed by variance, probability of occurrence, and/or a small spatial variance of precision.
- *Credibility,* also known as believability, comprises the content of evidence captured be a sensor which includes *veracity*, *objectivity*, *observational sensitivity*, and *self-confidence.*

Reliability from the engineering design domain (e.g., mean time between failures) refers to consistent ability to perform a function, and reliability of a source means consistently measuring the target phenomenon. It may be useful to model source failures over time using an exponential or Poisson distribution. For information fusion and systems analysis, we need both a source element (reliability) as well as a content element (credibility) to characterize information quality. Next, we describe the information model that consists of data sources from human and machines that requires uncertainty analysis.

## III. Information Fusion

### A. Information Fusion Evaluation

Information fusion combines information from multiple sources, distributions [32], or information over various system-level model processing levels as described in the *Data Fusion Information Group* (DFIG) model [33, 34, 35], depicted in Figure 1. The DFIG model outlines various processes for information fusion such as object assessment [36] (Level 1 – L1), situational assessment (L2), impact assessment (L3), and resource management (L4). Data and information fusion can be applied to assess the operating performance of algorithms [37], sources (reliability), as well as message content (credibility). For system-level analysis, it is important to look at source context reliability of humans (L5) and data sources for sensor (L4) and mission management (L6).

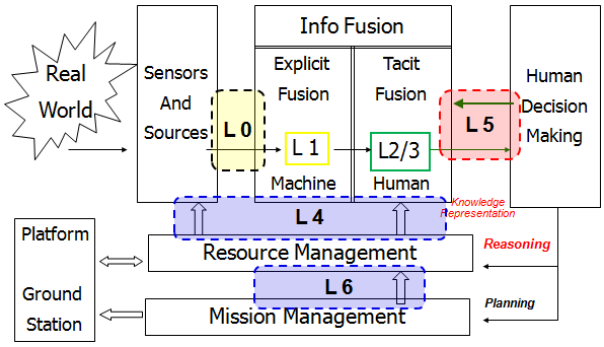


Figure 1 - DFIG Information Fusion model.

In the DFIG model, the goal is to separate information fusion (L0-L3) from sensor control, platform placement, and user selection to meet mission objectives (L4-L6) [38, 39, 40]. Information fusion across all the levels includes many metrics that need to be evaluated over uncertainty measures [41]. Challenges for information fusion, both at the hardware (i.e. components and sensors) and the software (i.e. algorithms and processes) levels were addressed by the *ETURWG* [http://eturwg.c4i.gmu.edu] [2]. Definitions of uncertainty measures such as accuracy [42], precision [43], reliability, and credibility are important for measures of effectiveness including validity and verification [44]. For example, accuracy (i.e., validity) measures distance from the truth, while precision (i.e., reliability) measures repeatability of results.

Examples of information fusion include tracking accuracy [45, 46], tracking filter credibility [47], and object detection credibility [48, 49] which are important for information quality and quality of service metrics [50].

### B. NATO STANAG 2511

For STANAG 2511, as an update to STANAG2022, there are general listings of categories for reliability and credibility that are of interest to the ETRUWG [51, 52, 53]. Table 1 lists the STANAG 2511 issues that provided initial discussion for the ETURWG and the subsequent discussions in the URREF. Reliability and credibility are independent criteria for evaluation. The resultant rating will be expressed in the appropriate combination of letter and number (STANAG 2511). Thus information received from a "usually reliable" source which is adjusted as "probably true" will be rated as "B2". Information from the same source of which the "truth cannot be judged" will be rated as "B6".

The URREF ontology, shown in Figure 2, distinguishes between reliability and credibility in evidence handling and evidence processing; respectively. In this paper, we utilize the STANAG 2511 definitions of reliability (of source) and credibility (of information). From the ETURWG discussions, credibility and reliability also relate to weight of evidence, relevance, and completeness; although others are currently being explored.

Table 1: STANAG 2511 Reliability and Credibility Relations and Definitions

| RELIABILITY | CODE | EXPLANATION From STANAG 2511 |
|---|---|---|
| Completely Reliable | A | A tried and trusted source which can be depended upon with *confidence* |
| Usually Reliable | B | A past successful source for which there is still some element of doubt in particular cases |
| Fairly Reliable | C | A past occasionally used source upon which some degree of confidence can be based |
| Not Usually Reliable | D | A source which has been used in the past but has proved more often than not unreliable |
| Unreliable | E | A source which has been used in the past and has proved unworthy of any confidence |
| Cannot be judged | F | It refers to a source which has not been used in the past |

| CREDIBILITY | CODE | EXPLANATION From STANAG 2511 |
|---|---|---|
| Confirmed | 1 | If it can be stated with *certainty* that the reported information originates from another source than the already existing information on the same object |
| Probably true | 2 | If the independence of the source cannot be guaranteed, but if, from the quantity and quality of previous reports, its *likelihood* is nevertheless regarded as sufficiently established |
| Possibly true | 3 | If insufficient confirmation to establish any higher degree of likelihood, a freshly reported item of information *does not conflict* with the previously reported target behavior |
| Doubtful | 4 | An item of information which tends to *conflict* with the previously reported or establish behavior pattern of an intelligence target |
| Improbable | 5 | An item of information which positively *contradicts* previously reported information of conflicts with the established behavior pattern of an intelligence target in a marked degree |
| Cannot be judged | 6 | If its truth cannot be judged |

## IV. Weight of Evidence

Weight of evidence (WOE) has different meanings in different contexts. A commonality is the need to integrate different sources or lines of evidence to form a conclusion or a decision.

In the field of *risk analysis*, WOE consists of a set of methods developed to assess the level of risks associated to factors or causes [54]. In most cases, WOE is a means of synthesizing information, while the solution adopted for weighing evidence is not explicit, or the evidence is presented without any interpretation. While some approaches rely on scoring techniques (see for instance research on sediments assessment described in [55]), the overall solutions remain qualitative in nature, developed for particular applications and poorly adaptable. Further discussion on WOE, as tackled within the risk analysis area is provided in [56].

WOE is addressed in a similar way in the *medical domain*, in relation to the rise of a new set of medical practices known as "evidence based medicine", promoting clinical solutions supported by practical experience, for which scientific support is not (yet) available.

From a different perspective, WOE is used in the *law and policy domain* to convey a subjective assessment of an expert analyzing different items of evidence, most often in relation to a causal hypothesis [57]. Intuitively, the concept is used to signify that the value of evidence must be above a critical threshold to support decisions or conclusions. In *law*, standards of evidence are recognized (for instance a three-level standard classifies evidence as "preponderance", "clear and convincing" and "beyond a reasonable doubt"), but experts will still exercise their judgment on the strength of evidence, as there is no methodology to assess this parameter. Without such methodologies, the variance in expert's judgments could be important, as subjective factors shape inevitably the outcome of the evidence in the evaluation.

### A. Weight of evidence for information fusion

In the field of information fusion, WOE captures the intuition

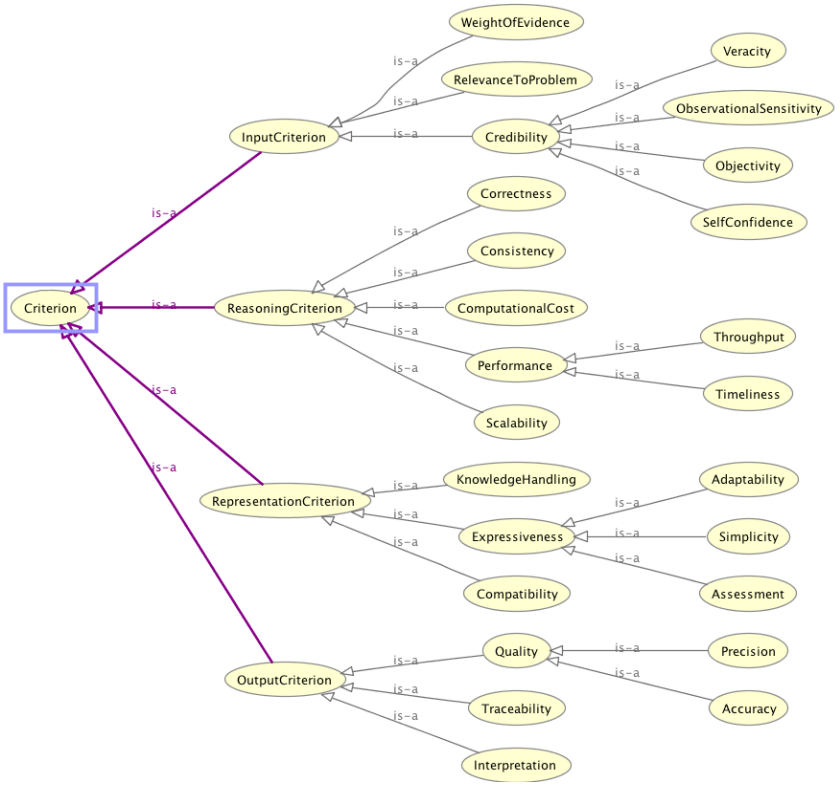


Figure 2 – URREF Ontology: Criteria Class [2].

that there is more or less evidence in the data, and this can be related to different parameters: the value of information itself (whether a piece of evidence conveys rich or poor information), the credibility of this information, in conjunction with the reliability of its source (can or should we believe this information), and finally the utility (or completeness) of this information with respect to a considered goal or task (is this data adding any detail to our existent data set?). WOE is an *attribute of information* and its values should be assessed by following a justifiable, repeatable and commonly accepted process. Therefore, several solutions have been developed to propose assessment mechanisms.

Among them, [58] proposes a probabilistic approach for information fusion where data items are weighted with respect to the accuracy or reliability of their source. This solution considers only independent information items and its adaptation to correlated information was developed [59]. In the field of evidential reasoning, the discounting operation introduced by Shafer [60], allows us to consider knowledge about the reliability of information sources. Smets and colleagues propose a method for learning a sensor's reliability, at various detail levels defined by users [61]. This method is generalized in Mercier, *et. al*. [62] by introducing the contextual discounting.

From a different perspective, [63] extends this frame in order to combine sources having different reliabilities and importance levels, while making a clear distinction between those notions.

It should be noticed that all references above consider only attributes of sources, while the weight of evidence should also be a function of information credibility. Underlying the same intuition of assigning different importance levels to items when fusing information, we can also cite research on prioritized and weighted aggregation operators, described in [64] and [65].

### *B. Relevance in Information Fusion*

Relevance has these components: property relation and piece of evidence (POE). Relevance is often considered as a relation between one property (or feature) and a conditional. That means that a property is relevant (or related) to another one "if it leads us to change our mind concerning whether the second property holds" [66].

For instance, in classification, relevance criteria determine how well a feature (a property) discriminates between the classes (another property). In this case, the feature selection step aims at identifying the features that are most relevant to the classification problem. We distinguish between the filter mode and the wrapping mode. In the filter mode, measures of relevance are used to characterize the features. In the wrapping mode, a classifier is used and the optimal subset of relevant features is the one which maximizes the given performance measures, such as the recognition rate, the area under curve, etc., subject to a penalty on the number of features. Classical relevance measures are based on: mutual information, distances between probabilities, cardinality distances, etc.

A *piece of evidence* (POE) is relevant if it impacts previous beliefs. In this case, the relevance of a piece of information can only be evaluated in conjunction with the combination (updating, revision) operator used, as the null element and the properties in general may differ from one operator to another. For example, in Information Retrieval, the process is used to assess the relevance of retrieved items (documents) based on a given query.

Measures of relevance are based on traditional recall and precision measures: Precision is the fraction of retrieved items that are relevant, and Recall is the fraction of relevant items that have been retrieved [67].

Relevance is defined with respect to a goal (or a context) and assesses quantitative and qualitative information change.

- *Quantitative approaches*: In quantitative approaches, the notion of relevance is often intimately linked to the notion of independence. For instance, in classical probability theory, according to Gärdenfors [68], a proposition $p$ is relevant to another proposition $r$ on evidence $e$ if $p$ and $r$ are conditionally dependent given $e$.
- *Qualitative approaches*: In qualitative approaches, the notion of relevance is linked to the material implication (see for instance the work of Goodman [69]): If $a$ then $b$, $a \rightarrow b$, then $a$ should be relevant to $b$.

### *C. How to evaluate a Relevance Criterion?*

First, we should clarify what is the object under evaluation, or what do we mean by uncertainty representation (UR). We follow here the distinction put forward in [70] about the difference between uncertainty *calculi* and decision procedures.

If UR means uncertainty *calculus* (UC) (mathematical framework, theory), then we are asking if, for instance, possibility theory or probability theory is able "*to capture how a given input is relevant* [...]", and to what degree. Although this is a very general question with certainly no binary answer, some evaluation could be done.

For instance, using a literature survey for document retrieval, what is needed is a notional scale. An example of a scale to be defined over methods, measures, or models :

A. exist and are well developed with the theory and results are significant;
B. exist but some further developments are required or results are not significant;
C. are missing, or
D. the concept is not addressed.

We could conclude for instance probability theory is very good at dealing with relevance since a plethora of methods and measures are defined (A), compared to possibility theory for which only few methods exist (B). This would be an *empirical evaluation*, mainly based on a literature survey. Although we could conclude that a theory is very good at dealing with the relevance concept (numerous methods, measures, papers etc), an absence of evidence in this sense for another theory would not mean that the latter is not good. Rather it would identify a research gap.

Each of the following elements can be evaluated separately:

(UC-1) The mathematical model for uncertainty representation

(UC-2) The uncertainty measures
(UC-3) The inference rules and combination operator
(UC-4) Transformation functions

If UR is a *decision procedure* (DP), we are asking if a particular algorithm, relying on possibly several theories, is able "*to capture how a given input is relevant* [...]", and to what degree. A DP distinguishes between the method and its implementation (e.g., fusion algorithm). Also, note that the same DP could be represented by several algorithms.

Two steps underlying may be distinguished:

1. Identification and assessment of pieces of information (or properties) according to their relevance; and
2. Filtering of irrelevant pieces of information.

Example of an experiment to be elaborated could be:

i. Consider a dataset with both relevant and irrelevant pieces of information;
ii. Each piece of information should have been previously labeled as relevant or irrelevant, possibly with some degrees;
iii. Run the decision procedure (fusion algorithm) with only relevant pieces of information and add progressively irrelevant (or less relevant) ones; and
iv. Evaluate the decision procedure based on other independent criteria such as the execution time, true positive rate, conclusiveness, interpretation, etc.

We could observe for instance that a given Decision Procedure, say DP-A, is better than another one, say DP-B, because its execution time is lower with an equivalent true positive rate. Even if DP-A is based on evidence theory and DP-B is based on probability theory, concluding that evidence theory is better for dealing with relevance than probability theory is obviously not trivial and would require special care.

A thinner-grained assessment of relevance criterion can be performed by assessing separately each of the following elements of an Atomic Decision Procedure (ADP):

(ADP-1) Universe of discourse
(ADP-2) Instantiated uncertainty representation
(ADP-3) Reasoning step
(ADP-4) Decision step

For instance, one could assess if one particular universe of discourse better allows expressing *relevance* concepts than another. Relevance contributes to WOE. Evaluating whether a representation is able to deal with relevance should rely on other criteria of the ontology (if UR is a decision procedure) and or on other empirical criteria to be defined (if UR is an uncertainty calculus). In addition to relevance affecting reliability and credibility, completeness needs to be considered.

*D. Evidence Completeness*

Reliability versus credibility is highly related to completeness of evidence. For example, we cannot postulate that: (P1) reliability of a source => credibility of information (that is more a source is reliable, more the credibility of the information it provides is high) WITHOUT assuming the completeness of pieces of evidences available for the source.

For example: (*Ming vase*): Let's consider an apparent Ming vase (a counterfeit or a genuine one) to be analyzed. Suppose that an expert provides his report based on only two attributes (say the shape and color of the vase) and concludes (based on these two attributes/pieces of evidences only) that the vase is a genuine Ming vase. Because it is based on this knowledge only, and because both attributes fit perfectly with those of a genuine Ming vase, the Expert is 100% reliable (he didn't make a mistake) in assessing the two attributes; however, we are still unsure of his reliability in assessing whether the vase is genuine. Additional POE if available may be 100% reliable and support the opposite conclusion. For example, let's suppose that when looking at the vase we see the printed inscription "Made in Taiwan". So we are now sure that we are facing a counterfeit Ming vase.

So we see that the reliability and credibility notions are highly dependent on the underlying completeness of pieces of evidence and the relationship of the evidence to the conclusion of interest. In the Ming vase example, if we treat the two attributes (color and shape) as complete evidence sufficient to establish the absolute truth, then if Expert is fully reliable, the information he/she provides becomes highly credible due to reliability of the source and completeness of the evidence.

When there is incompleteness of POE, nothing conclusive can be inferred about credibility unless some additional assumptions are introduced about the evidence necessary to establish the truth.

The fundamental question behind this, is to know if a source based only on local/limited knowledge (evidences) can (or not) conclude with an absolute certainty about an hypothesis, or its contrary so that any other/additional pieces of evidences cannot revise his/her conclusion. Depending on the standpoint we choose, we accept or reject (P1) which makes a big difference in reasoning. In summary, the ETURWG analysis highlights uncertainty elements of a WOE.

*E. URREF Weight of evidence*

With respect to criteria defined by URREF we can define weight of evidence as:

$$\text{WOE} = f(\text{Reliability, Credibility, Relevance, Completeness})$$

where $f$ is an function to be defined and relevance is related to the problem (or mission).

This is a translation of the following reasoning:

If (the source is reliable) then
If (the information provided is credible) then
If (this information is relevant to my problem) then
If (this information can enrich my existent information set) then
this information has some weight of evidence.

The four terms above are URREF criteria, while the last corresponds to a task-specific parameter that affects utility. For instance, utility can be evaluated by taking into consideration a distance between the set of information already available and a new item to determine utility completeness. Next, we demonstrate a modeling technique that brings together reliability and credibility to instantiate WOE calculations.

## V. Reliability and Credibility Analysis

A reliability assessment affects modern equipment systems performance capability, maintainability, usability, and the operational support cost. Knowing the system's reliability is important for efficient and effective performance. Due to the high complexity of system's engineering integration, it is difficult to evaluate system-level reliability. Some ways to estimate system-level reliability include: (1) predicting operational reliability based on design data, (2) statistically analyze operational data, or (3) develop performance models based on real-world operational constraints.

*Reliability* prediction depends on models, such as life-cycle analysis. Typical models include Poisson, Exponential, Weibull, or Bernoulli distributions. Standard components, operating for a long time, may have data to support a priori analysis and modeling; however, the likelihood of reliability effectiveness is subject to real-world conditions that have not been modeled. For exponentially distributed failure times, the density function and the cumulative distribution function for time to failure of the system components are:

$$f(t) = \lambda e^{-\lambda t} \quad ; \quad F(t) = 1 - e^{-\lambda t} \qquad (2)$$

The physical meaning of $F(t)$ is the probability that a failure (doubt) occurs *before* the time $t$ and $f(t)$ is the failure density: the probability that the component will fail in a small interval $t \pm \Delta t$ is given by $2f(t)\Delta t$. As $t$ increases, the value of $F(t)$ approaches 1 at $t = \infty$.

For a fusion or reliability metric of a source, we need to map the semantics into quantifiable metrics based on the source context. Here we assume that we take discrete measurement and a consistent source has almost no failures. On the other hand, a non consistent source fails quickly. As a quick look we show a notional example, but realize that for human sources this model does not hold. For example, to ascertain a "not usual source" is difficult to quantify and caution and improvements would be forthcoming from the ETURWG.

Classification systems process evidence features by an algorithm to classify evidence into classes. Results are tested against truth and reported using a confusion matrix (CM) [27]. A CM can thus be used to measure reliability of a classification system. A CM is an estimate of likelihoods of the accumulated evidence of classifier. The elements of a confusion matrix are $c_{ij} = \Pr\{\text{Classifier decides } o_j \text{ when } o_i \text{ is true}\}$, where $i$ is the true object class, $j$ is the assigned object class, and $i = 1, \ldots, N$ for $N$ true classes. The CM elements can be represented as probabilities as $c_{ij} = \Pr\{z = j \mid o_i\} = p\{z_j \mid o_i\}$. To determine an object declaration, we need to use Bayes' rule to obtain $p\{o_i \mid z_j\}$ which requires the class priors, $p\{o_i\}$. We denote the priors and likelihoods as column vectors

$$p(\overline{o}) = \begin{bmatrix} p(o_1) \\ p(o_2) \\ \vdots \\ p(o_N) \end{bmatrix} \; ; \; p(z_j \mid \overline{o}) = \begin{bmatrix} p(z_j \mid o_1) \\ p(z_j \mid o_2) \\ \vdots \\ p(z_j \mid o_N) \end{bmatrix}. \qquad (3)$$

For $M$ decisions, a confusion matrix would be of the form

$$C = \begin{bmatrix} p(z_1 \mid o_1) & p(z_2 \mid o_1) & .. & p(z_M \mid o_1) \\ p(z_1 \mid o_2) & p(z_2 \mid o_2) & .. & p(z_M \mid o_2) \\ \ldots & \ldots & \ddots & \ldots \\ p(z_1 \mid o_N) & p(z_2 \mid o_N) & .. & p(z_M \mid o_N) \end{bmatrix}. \qquad (4)$$

## VI. Results

For the simulation, we do both reliability and credibility assessment formulation to model the STANAG2511 criteria for uncertainty representation. Note that we assume completeness and relevance in these simulations.

### A. Reliability

For source reliability, the parameter of choice is λ, which captures the rate of time between failures. Figure 3 demonstrates the intuition that reliable and unreliable sources remain unreliable and reliable. However, the interesting cases are those which are termed "usually reliable" (code B) which affects the uncertainty analysis.

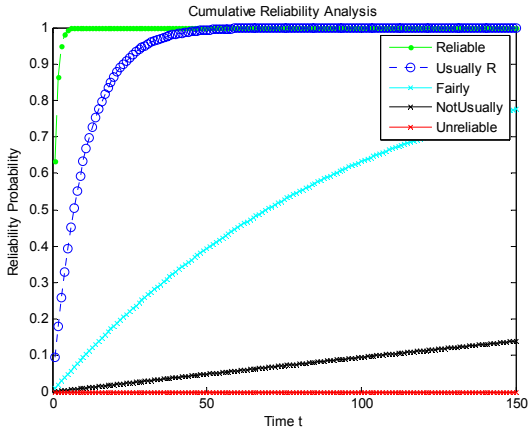


Figure 3 – Reliability Analysis

For Figure 3, a representative analysis of the reliability parameters are:

| | | |
|---|---|---|
| Code A | Completely Reliable : | λ = 0 |
| Code B | Usually Reliable : | λ = 0.001 |
| Code C | Fairly Reliable : | λ = 0.01 |
| Code D | Not Usually Reliable : | λ = 0.1 |
| Code E | Unreliable : | λ = 1 |
| Code F | Cannot be judged | λ undefined |

### C. Credibility

For credibility, since STANAG 2511 definitions deal with conflicts, we utilize comparisons between Dempster-Shafer Theory and the PCR5 rule. Setting up the modeling using CM of classifiers from the information content, we can develop representative CMs for the different definitions:

```
%%% Confusion Matrices for Classifiers (two sources)
CM1=[0.999 0.001; 0.001 0.999]
CM2=[0.95 0.05; 0.05 0.95]
CM3=[0.70 0.30; 0.30 0.70]
```

Now, we define credibility levels as follows, based on the confusion matrices of the two classifiers and whether or not their outputs agree:

| | |
|---|---|
| *Confirmed*: | CM1, outputs agree |
| *Probably* (independently confirmed): | CM2, outputs agree |
| *Possibly* (does not conflict): | CM3, outputs agree |
| *Doubtful* (tends to conflict): | CM3, outputs disagree |
| *Improbable* (conflicts): | CM1 or CM2, outputs disagree |

Figure 4 shows a comparison of the CM results of a "possibly true" (code 3) to validate that the PCR5 rule better supports evidence analysis than the Dempster-Shafer method.

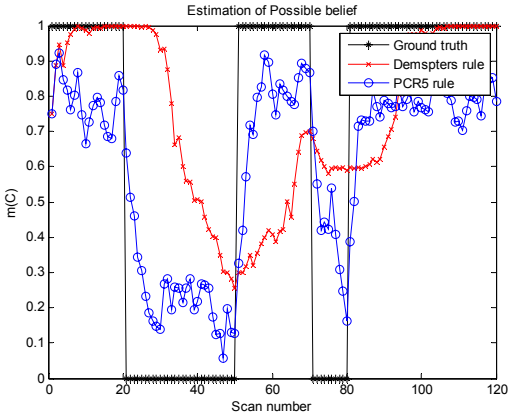


Figure 4 – DS versus PCR5 for "Possibly True" (Code 3)

Figure 5 and 6 highlight the credibility relations associated with a DS and PCR5 formulation, where PCR5 better represents an expected analysis for calculating the STANAG 2511 credibility codes.

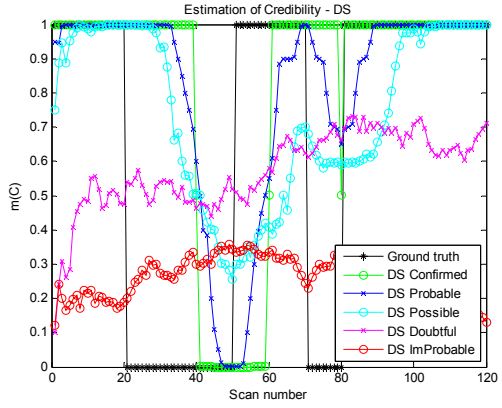


Figure 5 –DS Credibility of STANAG 2511 (Codes 1-5)

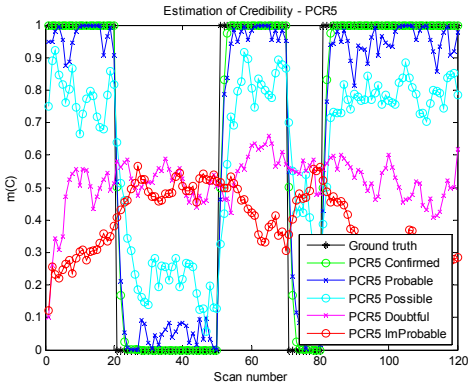


Figure 6 – PCR5 Credibility of STANAG 2511 (Codes 1-5)

## VII. Conclusions

In this paper, we overviewed uncertainty representation discussions from the ETURWG as related to the STANG 2511 reliability and credibility. In our URREF model for weight of evidence, included are relevance and completeness. We demonstrated modeling for reliability and credibility and provided simulations as related to evidence reasoning methods of the PCR5 rule. These results provide a more tractable (and mathematical) ability to calculate the STANAG 2511 codes.

Reliability and credibility affect higher levels of information fusion (i.e. beyond Level 2 fusion) grand challenges [71] of uncertainty representation [72], ontologies [73, 74] and uncertainty evaluation [75, 76]. Future research will further explore the uncertainty ontology within the URREF, use cases of real systems for a combined credibility/reliability assessment, and mathematical inclusion of other metrics such as relevance and completeness.

## References


[1] STANAG 2511 (January 2003, Intelligence reports, NATO Unclassified).

[2] P. C. G. Costa, K. B. Laskey, E. Blasch and A-L. Jousselme, "Towards Unbiased Evaluation of Uncertainty Reasoning: The URREF Ontology," *Int. Conf. on Info Fusion*, 2012.

[3] E. P. Blasch, E. Bosse, and D. Lambert, *High-Level Information Fusion Management and Systems Design*, Artech House, Norwood, MA, 2012.

[4] A. Peiravi, "Reliability Prediction of Electronic Navigation and Guidance Employing High Quality Parts to Achieve Increased Reliability," *J. Of App. Sciences*, 9 (16), 2009.

[5] R. Pan, "A Bayes Approach to Reliability Prediction Utilizing Data from Accelerated Life Tests and Field Failure Observations," *Quality and Reliability Engineering, International.* 25(2): 229-240, 2009.

[6] L. Bai, S. Biswas, and E. P. Blasch, "Survivability – An Information Fusion Process Metric From an Operational Perspective," *Int. Conf. on Info Fusion*, 2007.

[7] E. P. Blasch, M. Pribilski, B. Roscoe, *et. al.*, "Fusion metrics for dynamic situation analysis," *Proc. of SPIE,* Vol. 5429, 2004.

[8] E. Blasch, "Sensor, User, Mission (SUM) Resource Management and their interaction with Level 2/3 fusion" *Int. Conf. on Info Fusion*, 2006.

[9] E. P. Blasch, "Derivation of a Reliability Metric for Fused Data Decision Making," *IEEE NAECON Conf.*, 2008.

[10] L. Bai and E. P. Blasch, "Two-Way Handshaking Circular Sequential k-out-of-n Congestion System," *IEEE Trans. on Reliability*, Vol. 57, No. 1, pp. 59-70, Mar. 2008.

[11] E. Blasch, *Derivation of A Belief Filter for High Range Resolution Radar Simultaneous Target Tracking and Identification*, Ph.D. Dissertation, Wright State University, 1999.

[12] Y. Wu, J. Wang, J. Cheng, H. Lu, E. Blasch, L. Bai, and H. Ling, "Real-Time Probabilistic Covariance Tracking with Efficient Model Update," *IEEE Trans. on Image Processing*, **21**(5):2824-2837, 2012.

[13] C. Yang, L. Kaplan, and E. Blasch, "Performance Measures of Covariance and Information Matrices in Resource Management for Target State Estimation," *IEEE Trans. on Aerospace and Electronics*, Vol. 48, No. 3, pp. 2594 – 2613, 2012.

[14] E. Blasch, I. Kadar, K. Hintz, *et. al.*, "Resource Management Coordination with Level 2/3 Fusion Issues and Challenges," *IEEE Aerospace and Elect. Sys. Mag.*, Vol. 23, No. 3, pp. 32-46, Mar. 2008.

[15] M. Šimandl, M., and I. Punčochář, "Active-Fault detection and control: Unified formulation and optimal design," Vol. 45, *Automatica*, 2009.

[16] C. S. Byinton and A. Garga, "Data Fusion for Developing Predictive Diagnostics for Electromechanical Systems," Ch 23. in *Handbook of Data Fusion*, D. Hall and J Llinas (Eds.), CRC Press, 2001.

[17] E. Blasch, I. Kadar, J. Salerno, M. M. Kokar, S. Das, *et. al.*, "Issues and Challenges in Situation Assessment (Level 2 Fusion)," *J. of Advances in Information Fusion,* Vol. 1, No. 2, pp. 122 - 139, Dec. 2006.

[18] J. Llinas, "Assessing the Performance of Multisensor Fusion Processes," Ch 20 in *Handbook of Multisensor Data Fusion*, D. Hall and J. Llinas (Eds.), CRC Press, 2001.

[19] J. N. Yoo and G. Smith, "Reliability modeling for systems requiring mission reconfigurability," *Proc. Reliability and Maintainability Symp.*, pp. 133–139, 1990.

[20] G. Seetharaman, A. Lakhotia, E. Blasch, "Unmanned Vehicles Come of Age: The DARPA Grand Challenge," *IEEE Computer Society Magazine*, Vol. 39, No. 12, pp. 26-29, Dec 2006.

[21] K. M. Lee, Z. Zhi, R. Blenis, and E. P. Blasch, "Real-time vision-based tracking control of an unmanned vehicle," *Journal of Mechatronics - Intelligent Motion Control*, Vol. 5, No. 8, pp. 973 –991, 1995.

[22] T. Bass and R. Robichaux, "Defense – In-Depth Revisited: Qualitative Risk Analysis Methodology for Complex Network-centric operations," *IEEE MILCOM*, 2001.

[23] J. Salerno, E. Blasch, M. Hinman, and D. Boulware, "Evaluating algorithmic techniques in supporting situation awareness," *Proc. of SPIE,* Vol. 5813, April 2005.

[24] G. Chen, D. Shen, C. Kwan, J. Cruz, M. Kruger, and E. Blasch, "Game Theoretic Approach to Threat Prediction and Situation Awareness," *J. of Advances in Information Fusion,* Vol. 2, No. 1, 1-14, June 2007.

[25] G. Chen, E. P. Blasch, and L. Haynes, "A game Theoretic Data Fusion Approach for Cyber Situational Awareness," *Cyber Fusion Conf*., 2007.

[26] D. Shen, G. Chen, J. Cruz, et al., "Game Theoretic Solutions to Cyber Attack and Network Defense Problems," *ICCRTS*, Nov. 2007.

[27] B. Kahler and E. Blasch, "Decision-Level Fusion Performance Improvement from Enhanced HRR Radar Clutter Suppression," *J. of. Advances in Information Fusion,* Vol. 6, No. 2, Dec. 2011.

[28] J. Patel, *A Trust and Reputation Model for Agent-based Virtual Organizations*, PhD Univ. Southampton, 2007.

[29] J. Dezert, A. Tchamova, F. Smarandache, and P. Konstantinova, "Target Type Tracking with PCR5 and Dempster's Rules: a comparative analysis," *Int. Conf. on Information Fusion*, 2006.

[30] A. Martin, C. Osswald, J. Dezert, and F. Smarandache, "General Combination Rules for Qualitative and Quantitative Beliefs," *J. of Adv. in Information Fusion*, Vol. 3, No. 2, December, 2008.

[31] E. Blasch, J. Dezert, and P. Valin, "DSMT Applied to Seismic and Acoustic Sensor Fusion," *Proc. IEEE Nat. Aerospace Elect Conf*, 2011.

[32] E. P. Blasch and M. Hensel, "Fusion of Distributions for Radar Clutter modeling," *Int. Conf. on Info Fusion*, 2004.

[33] E. P. Blasch et al, "JDL Level 5 Fusion model 'user refinement' issues and applications in group Tracking," *Proc. SPIE,* Vol. 4729, 2002.

[34] E. P. Blasch, "Level 5 (User Refinement) issues supporting Information Fusion Management," *Int. Conf. on Info Fusion*, 2006.

[35] E. Blasch, *et. al*., "DFIG Level 5 (User Refinement) issues supporting Situational Assessment Reasoning," *Int. Conf. on Info Fusion,* 2005.

[36] C. Yang and E. Blasch, "Kalman Filtering with Nonlinear State Constraints," *IEEE Trans. Aero. and Elect. Systems*, Vol. 45, No. 1, 70-84, Jan. 2009.

[37] Z. Liu, E. Blasch, Z. Xue, R. Langaniere, and W. Wu, "Objective Assessment of Multiresolution Image Fusion Algorithms for Context Enhancement in Night Vision: A Comparative Survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, **34**(1):94-109. 2012.

[38] E. P. Blasch and P. Hanselman, "Information Fusion for Information Superiority," *IEEE Nat. Aerospace and Electronics Conference,* 2000.

[39] E. Blasch, "Situation, Impact, and User Refinement," *Proc. of SPIE,* Vol. 5096, April 2003.

[40] E. Blasch, "User refinement in Information Fusion", Chapter 19 in *Handbook of Multisensor Data Fusion 2nd Ed,* (Eds.) D. Hall, and J. Llinas, CRC Press, 2008.

[41] P. Hanselman, C. Lawrence, E. Fortunano, B. Tenney, and E. Blasch, "Dynamic Tactical Targeting," *Proc. of SPIE,* Vol. 5441, 2004.

[42] C. Yang and E. Blasch, "Pose Angular-Aiding for Maneuvering Target Tracking," *Int. Conf. on Info Fusion*, July 2005.

[43] E. P. Blasch, *et. al*., "Ontology Alignment using Relative Entropy for Semantic Uncertainty Analysis," *Proc. IEEE NAECON*, 2010.

[44] E. Blasch, P. Valin, E. Bossé, "Measures of Effectiveness for High-Level Fusion," *Int. Conference on Information Fusion*, 2010.

[45] X. Mei, H. Ling, Y. Wu, E. P. Blasch, and L. Bai, "Minimum Error Bounded Efficient L1 Tracker with Occlusion Detection," *IEEE Computer Vision and Pattern Recognition*, 2011.

[46] Y. Wu, E. Blasch, G. Chen, L. Bai, and H. Ling, "Multiple Source Data Fusion via Sparse Representation for Robust Visual Tracking," *Int. Conf. on Info Fusion*, 2011.

[47] E. Blasch, O. Straka, J. Duník, and M. Šimandl, "Multitarget Performance Analysis Using the Non-Credibility Index in the Nonlinear Estimation Framework (NEF) Toolbox," *Proc. IEEE Nat. Aerospace Electronics Conf (NAECON)*, 2010.

[48] E. Blasch, R. Breton, and P. Valin, "Information Fusion Measures of Effectiveness (MOE) for Decision Support," *Proc. SPIE* 8050, 2011.

[49] Y. Zheng, W. Dong, and E. Blasch, "Qualitative and quantitative comparisons of multispectral night vision colorization techniques," *Optical Engineering*, Vol. 51, Issues 8, Aug. 2012.

[50] C. Bisdikian, L. M. Kaplan, M. B. Srivastava, D. J. Thornley, D. Verma, and R. I. Young, "Building principles for a quality of information specification for sensor fusion," *Int. Conf. on Information Fusion*, 2009.

[51] V. Nimier, "Information Evaluation: a formalization of operational recommendations," *Int. Conf. on Information Fusion*, 2004.

[52] J. Besombes, and A. R. d'Allonnes, "An Extension of STANAG2022 for Information Scoring," *Int. Conf. on Info. Fusion*, 2008.

[53] E. P. Blasch, R. Breton, and P. Valin, "Information Fusion Measures of Effectiveness (MOE) for Decision Support," *Proc. SPIE* 8050, 2011.

[54] Linkov I, Welle P, Loney D, Tkachuk A, Canis L, Kim JB, Bridges T., "Use of multicriteria decision analysis to support weight of evidence evaluation," *Risk Anaysis*, August, 31(8):1211-25, 2011.

[55] P. Chapman, "A decision making framework for sediment assessment developed for the Great Lakes," *Human and Ecological Risk Assessment*, 8(7):1641–1655, 2002.

[56] D. L. Weed, "Weight of evidence: a review of concept and method." *Risk Analysis,* 25(6), 1545-1557, 2005.

[57] S. Krimsky, "The weight of scientific evidence in policy and law," *American Journal of Public Health,* 95 (S1), S129-S136, 2005.

[58] L. Peijun, and S. Benqin, "Land cover classification of multi-sensor images by decision fusion using weight of evidence model," International Archives of the Photogrammetry, *Remote Sensing and Spatial Information Sciences*, XXII ISPRS Congress, 2012.

[59] M. Deng, "A Conditional Dependence Adjusted Weights of Evidence Model. Natural Resources Research," 18(4), 249- 258, 2009.

[60] G. Shafer, *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.

[61] Z. Elouedi, K. Mellouli, and Ph. Smets, "Assessing sensor reliability for multisensor data fusion with the transferable belief model" *IEEE Tr. on Systems, Man and Cybernetics* B, volume 34, pages 782-787, 2004.

[62] D. Mercier, T. Denoeux, and M.H. Masson, "Refined sensor tuning in the belief function framework using contextual discounting," *Proc. of IPMU*, Vol II, pages 1443-1450, 2006.

[63] F. Smarandache, J. Dezert J.-M. Tacnet, "Fusion of sources of evidence with different importances and reliabilities, *Int. Conf. on Info Fusion* 2010.

[64] R. Yager, "Prioritized aggregation operators," *International Journal of Approximate Reasoning*, Vol. 48, Issue 1, pages 263-274, 2008.

[65] R. Yager, "Prioritized operators and their applications," *Proc. of the 6th IEEE International Conference 'Intelligent Systems'*, 2012.

[66] J. P. Delgrande, and F. J. Pelletier, "A Formal Analysis of Relevance," *Erkenntnis*, 49(2):137–173, 1998.

[67] P. Borlund, "The Concept of Relevance in IR", *Journal of the American Society for Information Sciences and Technology*, 54(10):913–925, 2003.

[68] P. Gärdenfors, "On the logic of relevance," *Synthese*, 37:351–367, 1978.

[69] N. Goodman. About. *Mind*, 70:1–24, 1961.

[70] A.-L. Jousselme, and P. Maupin, "A brief survey of comparative elements for uncertainty calculi and decision procedures assessment", *Panel discussion, Int. Conf,. on Information Fusion*, 2012.

[71] E. P. Blasch, D. A. Lambert, P. Valin, *et. al*., "High Level Information Fusion (HLIF) Survey of Models, Issues, and Grand Challenges," *IEEE Aerospace and Elect. Sys. Mag.,* Vol. 27, No. 8, Aug. 2012.

[72] P.C.G. Costa, R.N. Carvalho, K.B. Laskey, and C.Y. Park, "Evaluating Uncertainty Representation and Reasoning in HLF systems," *Int. Conference on Information Fusion*, 2011.

[73] E. Blasch, "Ontological Issues in Higher Levels of Information Fusion: User Refinement of the Fusion Process," *Int. Conf. on Info Fusion,* 2003.

[74] P.C.G. Costa, KC Chang, K.B. Laskey, T. Levitt, and W. Sun, "High-Level Fusion: Issues in Developing a Formal Theory," *Int. Conf. on Information Fusion*, 2010.

[75] E. Blasch, P. C. G. Costa, K. B. Laskey, D. Stampouli, *et al*., "Issues of Uncertainty Analysis in High-Level Information Fusion – Fusion2012 Panel Discussion," *Int. Conf. on Info Fusion*, 2012.

[76] P. C. G. Costa, E. P. Blasch, K. B. Laskey, S. Andler, J. Dezert, A-L. Jousselme, and G. Powell, "Uncertainty Evaluation: Current Status and Major Challenges – Fusion2012 Panel Discussion," *Int. Conf. on Info Fusion*, 2012.