# Evaluating Precise and Imprecise State-Based Anomaly Detectors for Maritime Surveillance

**Christoffer Brax[1,2], Alexander Karlsson[1], Sten F. Andler[1], Ronnie Johansson[1], Lars Niklasson[1]**

Informatics Research Centre[1]
University of Skövde, Sweden
<first name>.<last name>@his.se

Electronic Defense Systems[2]
Saab AB, Sweden
<first name>.<last name>@saabgroup.com

***Abstract - We extend the State-Based Anomaly Detection approach by introducing precise and imprecise anomaly detectors using the Bayesian and credal combination operators, where evidences over time are combined into a joint evidence. We use imprecision in order to represent the sensitivity of the classification regarding an object being normal or anomalous. We evaluate the detectors on a real-world maritime dataset containing recorded AIS data and show that the anomaly detectors outperform previously proposed detectors based on Gaussian mixture models and kernel density estimators. We also show that our introduced anomaly detectors perform slightly better than the State-Based Anomaly Detection approach with a sliding window.***

**Keywords:** Anomaly detection, maritime surveillance, Bayesian combination operator, credal combination operator.

## 1 Introduction

*Maritime Domain Awareness* (MDA) is important for all maritime authorities and involve comprehension of all maritime activities and their possible impact on security, safety, environment and economy [1]. The objective is to anticipate threats such as terrorists, piracy and organized crime. According to the Department of Homeland Security [1], technological support for collecting, fusing and analyzing data is needed to identify trends and find anomalies. *Anomaly detection* is one of many enabling technologies for MDA. The goal of anomaly detection is to find "objects that are different from most other objects" [2]. The anomaly detection method shall "discover the real anomalies and avoid falsely labeling normal objects as anomalous" [2]. In essence, a well performing anomaly detection method should have a high detection rate without too many false alarms. Portnoy et al. [3] made the following definition in the domain of network intrusion detection:

"Anomaly detection approaches build models of normal data and then attempts to detect deviations from the normal model in observed data." (p.2)

Based on this definition one way of detecting anomalies is to use normal data to build a model that describes what is considered to be normal. This model can then be used to classify new data as normal or anomalous. Applied to maritime surveillance we want to capture normal behaviors of vessels in the normal model. The *State-Based Anomaly Detection* (SBAD) approach tries to solve this problem. There are however some aspects of the anomaly detection problem that needs further work. If the anomaly classification of an object is based on just one observation, we risk a significant amount of false alarms. To minimize the false alarms we could base the object classification on multiple observations over time. The problem however, is how one should perform such classification. In this paper we propose an extension of the original SBAD approach using the Bayesian and credal combination operators (cf [4; 5; 6]) to cope with this problem.

The paper is organized as follows; in Section 2 we present the SBAD approach and the Bayesian and credal combination operators. In Section 3 we introduce the precise and imprecise anomaly detectors which are based on the combination operators described in Section 2. In Section 4 we perform two experiments, the first one measure the detection delay and the second the precision and recall. We compare the result of our introduced detections with previous detectors. We conclude with a brief summary and conclusions.

## 2 Background

We describe the State-Based Anomaly Detection (SBAD) approach and how evidences can be combined into a joint evidence by using the Bayesian and credal combination operators.

### 2.1 State-Based Anomaly Detection

The SBAD approach [7; 8; 9; 10; 11] is based on a discrete state representation for both contextual and kinematic information. The representation is built upon a number of *atomic states* represented by *discrete random variables* $Y_1, \dots, Y_n$ all of which have a state space $\Omega_Y$. Let $y$ be any state in the state space $\Omega_Y$, i.e. $y \in \Omega_Y$. Each $Y_i$, $i \in \{1, \dots, n\}$, represents a *discretised feature* from the kinematic or contextual information. As an example, $Y_i$

can denote the course of the object of interest, where the state space is $\Omega_Y \triangleq \{north, south, east, west\}$. The state space $\Omega_{Y_i}$ for each $Y_i$ is set by a domain expert based on what can be expected in the dataset from the given domain. This way *expert knowledge* can be included into the representation. The discrete state spaces allow for a higher level representation of continuous features. Atomic states $Y_i$, $i \in \{1, \ldots, n\}$ can be combined into a *composite state* $\mathbb{Y} \triangleq Y_1 \times \ldots \times Y_n$ that captures the dependencies between atomic states. For example, stating that a vessel travel north is not very informative if we do not include the position. Stating that a vessel travel north on a certain position can be much more informative.

By utilizing a *training dataset*, consisting of a number of tracks where each track is a sequence of composite states $\langle \mathbb{y}_1, \ldots, \mathbb{y}_t \rangle \in \Omega_{\mathbb{Y}_1 \times \ldots \times \mathbb{Y}_t}$, we can build normal models that capture different aspects of the domain of interest. In order to capture the normality with respect to composite states, we can build a normal model by utilizing the *relative frequency* of the *composite state occurrences* in the training dataset, i.e.:

$$p_{Occ}(\mathbb{y}) \triangleq \frac{n(\mathbb{y})}{\sum_{\mathbb{y} \in \Omega_{\mathbb{Y}}} n(\mathbb{y})} \quad (1)$$

where $\mathbb{y} \in \Omega_{\mathbb{Y}}$ and $n(\mathbb{y})$ denotes the number of observations of $\mathbb{y}$. Another aspect that can be interesting to capture by a normal model is *transitions* between composite states $\langle \mathbb{y}_i, \mathbb{y}_j \rangle \in \Omega_{\mathbb{Y} \times \mathbb{Y}}$:

$$p_{Tr}\big(\langle \mathbb{y}_i, \mathbb{y}_j \rangle\big) \triangleq \frac{n\big(\langle \mathbb{y}_i, \mathbb{y}_j \rangle\big)}{\sum_{\langle \mathbb{y}_i, \mathbb{y}_j \rangle \in \Omega_{\mathbb{Y} \times \mathbb{Y}}} n\big(\langle \mathbb{y}_i, \mathbb{y}_j \rangle\big)} \quad (2)$$

where $n\big(\langle \mathbb{y}_i, \mathbb{y}_j \rangle\big)$ denotes the number of transitions $\langle \mathbb{y}_i, \mathbb{y}_j \rangle$ over all tracks and where $i$ and $j$ denote two consecutive time steps from the same track.

We have previously used the above models as a basis for constructing binary anomaly detectors which classify each observation as anomalous or normal [8]. In order to detect anomalies that develops over time and decrease the risk of false alarms, we utilized a *sliding window* with a pre-determined size ($n_{ws}$) when classifying a track, i.e., a specific number of consecutive observations are considered. The number of anomalous classifications from each detector, within the sliding window, is fused by using a *weighted sum*. To classify a track as anomalous, the weighted sum must reach a specific user specified threshold ($T_{al}$). See [8] for more information about the SBAD approach with a sliding window.

The SBAD approach has previously been used for detecting behavioral anomalies in video surveillance data [8; 10; 11]. The method has also been used for detecting anomalies in maritime vessel behavior and for transportation security based on airborne radar information [7] and GPS information [9] respectively.

## 2.2 Bayesian Combination Operator

Assume that we are interested in a random variable $X$, with a corresponding state space $\Omega_X$, and where observations $y_1$ and $y_2$, are considered to be *evidence* for $X$. In Bayesian theory [12], such evidences are represented by *likelihood functions* $p(y_1|X)$ and $p(y_2|X)$. By assuming the observations $y_1$ and $y_2$ are *conditionally independent* given $X$, we can construct a *joint evidence* $p(y_1, y_2|X)$ by:

$$p(y_1, y_2|X) = p(y_1|X)p(y_2|X) \quad (3)$$

One problem with Equation (3) is that the joint evidence monotonically decreases with the number of combinations. Thus, when implementing such combination operator, one will eventually obtain erroneous results due to the limited precision of the double representation. For this reason it is convenient to formulate Equation (3) in an alternative way where one has *normalized* the likelihood functions and result of the combination. The *Bayesian combination operator* can now be defined as (cf [4; 5; 6]):

**Definition 1:** *The Bayesian combination operator* $\Phi_{\mathcal{B}}(\hat{p}(y_1|X),\ \hat{p}(y_2|X))$ *is defined as:*

$$\Phi_{\mathcal{B}}\big(\hat{p}(y_1|X), \hat{p}(y_2|X)\big) \triangleq \frac{\hat{p}(y_1|X)\hat{p}(y_2|X)}{\sum_{x \in \Omega_X} \hat{p}(y_1|x)\hat{p}(y_2|x)} \quad (4)$$

*where* $\hat{p}(y_i|X)$, $i \in \{1,2\}$, *are conditionally independent normalized likelihood functions. The operator is undefined when* $\sum_{x \in \Omega_X} \hat{p}(y_1|x)\hat{p}(y_2|x) = 0$.

## 2.3 Credal Combination Operator

One criticism of the $\Phi_{\mathcal{B}}$ operator is that it is necessary to specify the involved evidences, i.e., the likelihood functions, in a *precise* way. Such assumption is not always realistic, in particular for problems where only scarce information about the state of interest exists. For such type of problems it has been proposed [13] that one should utilize *imprecision* in probabilities as a way of model *lack of information.* Imprecision in probabilities induces a so called *credal set* [14; 15], i.e., *a closed convex set of probability functions.* In fact, if we allow imprecision in belief and evidences, i.e., we represent *belief* by a credal set and *evidence* by a *closed convex set of likelihood functions*, we have what is referred to as *credal set theory* (cf, [6; 15; 16] ). The straightforward generalization of $\Phi_{\mathcal{B}}$ operator would then be to allow imprecision in the evidences. Let $\hat{\mathcal{P}}(y|X)$ denote a closed convex set of normalized likelihood functions, i.e., a credal set. We first need to define our notion of

conditional independence for credal sets, denoted by *strong independence* [17]:

**Definition 2:** *Two convex sets of normalized likelihood functions are strongly conditionally independent iff all* $\hat{p}(y_1, y_2|X) \in \mathcal{E}(\hat{\mathcal{P}}(y_1, y_2|X))$ *can be expressed as* $\hat{p}(y_1, y_2|X) = \hat{p}(y_1|X)\hat{p}(y_2|X)$ *where* $\hat{p}(y_i|X) \in \hat{\mathcal{P}}(y_i|X), i \in \{1,2\}$, *and where* $\mathcal{E}(\cdot)$ *denotes the set of extreme points.*

The imprecise correspondence to the $\Phi_{\mathcal{B}}$ operator, namely, the *credal combination operator*[1] [4; 5; 6] can now be defined as:

**Definition 3:** *The credal combination operator* $\Phi_{\mathcal{C}}(\hat{\mathcal{P}}(y_1|X), \hat{\mathcal{P}}(y_2|X))$ *is defined as:*

$$\Phi_{\mathcal{C}}\left(\hat{\mathcal{P}}(y_1|X), \hat{\mathcal{P}}(y_2|X)\right) \triangleq \mathcal{CH}\left(\left\{\Phi_{\mathcal{B}}\left(\hat{p}(y_1|X), \hat{p}(y_2|X)\right) : \hat{p}(y_1|X) \in \hat{\mathcal{P}}(y_1|X), \hat{p}(y_2|X) \in \hat{\mathcal{P}}(y_2|X)\right\}\right) \quad (5)$$

*where* $\hat{\mathcal{P}}(y_i|X)$, $i \in \{1,2\}$, *are strongly conditional independent closed convex sets of normalized likelihood function, and where* $\mathcal{CH}(\cdot)$ *is the convex hull operator. The operator is undefined iff there exists* $\hat{\mathcal{P}}(y_i|X)$, $i \in \{1,2\}$ *such that* $\Phi_{\mathcal{B}}\left(\hat{p}(y_1|X), \hat{p}(y_2|X)\right)$ *is undefined.*

We see that the $\Phi_{\mathcal{C}}$ operator is a straightforward generalization of the $\Phi_{\mathcal{B}}$ operator. In order to be able to perform computation with the $\Phi_{\mathcal{C}}$ operator, one uses operand credal sets that are *polytopes,* i.e., sets that has a finite number of extreme points. The reason for this is that one can then compute the operator by using the *extreme points* of the operands [4; 5; 6].

## 3 Precise and Imprecise Anomaly Detectors

We elaborate on how the normal models from the SBAD approach (see Section 2.1) can be used in order to construct precise and imprecise anomaly detectors.

### 3.1 Precise Anomaly Detectors

Consider the case of where we receive a new observation $\mathbb{y}$ from an object and that we want to construct a corresponding evidence regarding the object being anomalous or not. Let $p(\mathbb{z})$ be any of $p_{Occ}(\mathbb{y})$ (see Equation 1) or $p_{Tr}(\langle \mathbb{y}_t, \mathbb{y}_{t+1} \rangle)$ (see Equation 2). Intuitively, the less probable with respect to the normal model $p(\mathbb{z})$ it is to observe a specific $\mathbb{z}$, the *stronger evidence* for anomaly. Let us introduce a random variable $X$ for an object being anomalous or normal, i.e., we have a state space $\Omega_X = \{a, n\}$ (anomaly or normal). Based on this line of reasoning, we can construct evidences $\hat{p}(\mathbb{z}|X)$ in the following way:

$$\hat{p}(\mathbb{z}|a) \triangleq \begin{cases} 0.5 + \dfrac{T - p(\mathbb{z})}{T} s_a, & p(\mathbb{z}) < T \\ 0.5 - \dfrac{p(\mathbb{z}) - T}{\left(\max\limits_{\mathbb{z} \in \Omega_{\mathbb{z}}} p(\mathbb{z})\right) - T} s_n, & p(\mathbb{z}) \geq T \end{cases} \quad (6)$$

$$\hat{p}(\mathbb{z}|n) \triangleq 1 - \hat{p}(\mathbb{z}|a)$$

where $s_a$ and $s_n$ are parameters that models the maximum possible strength that an evidence can constitute for anomaly respectively normality, and where $T$ is a threshold that constitutes the limit between normality and anomaly. The mapping in Equation (6) is seen in Figure 1.

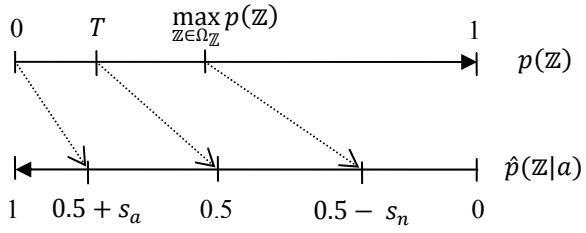


Figure 1: Mapping from $\mathrm{p}(\mathbb{z})$ to $\hat{\mathrm{p}}(\mathbb{z}|a)$.

We see that when we have not made any observations at all in the normal model, we have the strongest possible evidence for anomaly and the other way around for normality. It can be somewhat counterintuitive to think of an observation $\mathbb{y}$ as constituting evidence for normality. However, if one wants an object to be able to "recover" from an anomalous state, such modeling approach is necessary. Note that $s_n = 0$ implies that there exists no evidence for normality. The threshold $T$ is a parameter that can be set with respect to the application at hand in order to adjust the sensitivity of the detector.

Now, at each time step $t$ we obtain an observation $\mathbb{y}_t$, which we can use in Equation (6) for obtaining a corresponding evidence $\hat{p}(\mathbb{z}_t|X)$. In order to simplify notation, let us introduce:

$$\hat{p}_t(X) \triangleq \hat{p}(\mathbb{z}_t|X) \quad (7)$$

We can now use the $\Phi_{\mathcal{B}}$ operator in order to combine all evidences that has been obtained regarding a specific object into a joint evidence $\hat{p}_{0:t}(X)$, i.e.:

$$\hat{p}_{0:t}(X) \triangleq \Phi_{\mathcal{B}}\left(\ldots \Phi_{\mathcal{B}}\left(\hat{p}_0(X), \hat{p}_1(X)\right) \ldots, \hat{p}_t(X)\right) \quad (8)$$

[1] Arnborg introduced this operator as *"the robust Bayesian combination operator"*, however, we deliberately avoid this terminology since "*robust Bayesianism"* refers to a sensitivity interpretation of the imprecision and we do not want to exclude other possible interpretations, e.g. [13].

where $\hat{p}_0(X)$ denotes the *prior evidence* which is specified by the user.

## 3.2 Anomaly Classification

Let $\hat{p}(X)$ and $\hat{p}_{0:t}(X)$ denote evidence respectively joint evidence for any of the anomaly detectors that we have introduced. The question then is when a certain object should be classified as an anomaly? In principle, such classification can be performed by introducing a threshold $T' \in [0,1]$ and classify the object as anomalous when:

$$\hat{p}_{0:t}(a) \geq T' \tag{9}$$

Such classification schema is not only dependent on the threshold $T'$ but also on the maximum possible strengths $s_a$ and $s_n$. Consider for example an object that has generated a large number of normal observations followed by only a few anomalous observations. In such case the joint evidence before taking the anomalous observations into account will constitute strong evidence for normality. Whether or not one classify the object as an anomaly in such situation depends on all parameters $s_a$, $s_n$ and $T'$. For this reason, we propose that these parameters should be set jointly with respect to the application at hand by determining the number of *extreme anomaly evidences,* i.e., $\hat{p}(a) = 0.5 + s_a$, one require, starting from the prior evidence $\hat{p}_0(a)$, in order to classify the object as an anomaly. Furthermore, in order to guarantee a certain level of reactivity of the detector, it is necessary to limit the joint evidence $\hat{p}_{0:t}(n)$ for normality to a minimum that is equal to the prior evidence, i.e.:

$$\min_{\langle \mathbb{z}_1,\ldots,\mathbb{z}_t \rangle} \hat{p}_{0:t}(n) \geq \hat{p}_0(n) \tag{10}$$

for all tracks $\langle \mathbb{z}_1, \ldots, \mathbb{z}_t \rangle$. Such limitation ensures a certain minimum degree of reactivity of the anomaly detectors.

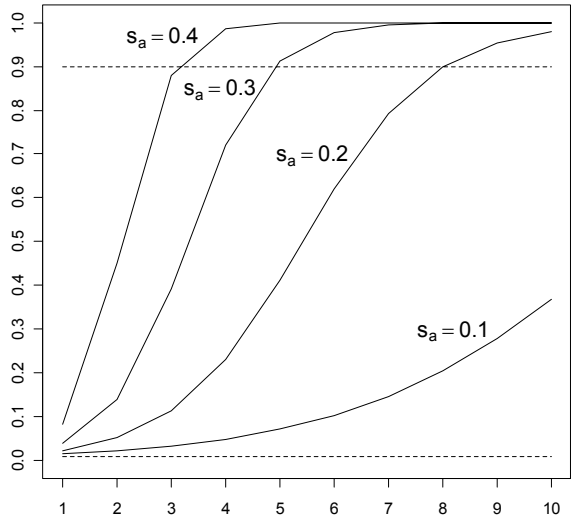


Figure 2: The figure depicts different choices of the parameter $s_a$ where $T' = 0.9$ (upper dahshed line). The x-axis shows the number of combinations and the y-axis shows the strength of the joint evidence $\hat{p}_{0:t}(X)$. The lower dashed line shows the prior evidence $\hat{p}_0(n)$.

An example of the procedure is seen in Figure 2. From the figure, we see that if we want an object to be classified as an anomaly after eight extreme observations, we need to set $s_a \approx 0.2$. When $s_a$ has been set by using our proposed method, one also need to consider the maximum strength of an evidence for normality, i.e., $s_n$. The question that needs to be addressed when choosing such strength is to what extent one wants a normal observation to influence the joint evidence towards normality. This can be modeled by considering the ratio between strengths $s_a$ and $s_n$.

## 3.3 Imprecise Anomaly Detectors

One problem with the anomaly detectors that we introduced in the previous section is that the threshold $T$ needs to be specified precisely. Such *precise thresholds* constitute a sharp border of what observations that constitute evidence for anomaly or not, and one may not be comfortable with such sharpness. Moreover, since evidences are combined over time, small variations in the thresholds can potentially have a significant impact on the joint evidence, which in the end can lead to different classifications of an object. For this reason, let us adopt an *interval of thresholds* instead, i.e., we allow for *imprecision* in the threshold. Let us denote the lower and upper bound in such interval for $\underline{T}$ and $\overline{T}$. Consider Figure 1, which shows the mapping between an observation and the corresponding evidence. Now instead of using the single threshold $T$, let us use each threshold in the interval $[\underline{T}, \overline{T}]$ for obtaining evidences. From the figure, we see that $[\underline{T}, \overline{T}]$ yields a set of evidences that is *closed and convex*, i.e., a *credal set* $\hat{\mathcal{P}}(X)$ such that:

$$\hat{\mathcal{P}}(X) = \{\hat{p}(X) \; as \; in \; Equation \; (6) : \; T \in [\underline{T}, \overline{T}]\} \tag{11}$$

Now by using the $\Phi_{\mathcal{C}}$ operator, we can obtain the joint credal evidence by:

$$\hat{\mathcal{P}}_{0:t}(X) \triangleq \Phi_{\mathcal{C}}\big(\ldots \Phi_{\mathcal{C}}\big(\hat{\mathcal{P}}_0(X), \hat{\mathcal{P}}_1(X)\big) \ldots, \hat{\mathcal{P}}_t(X)\big) \tag{12}$$

Note that since the credal operands in Equation (12) are intervals ($|\Omega_X| = 2$), i.e., there are only two extreme points, computation of the $\Phi_{\mathcal{C}}$ operator is easily performed.

One problem that arises when using imprecise anomaly detectors is that it is not as clear how one should classify an object due to the fact that:

$$\underline{\hat{\mathcal{P}}_{0:t}}(a) < T' < \overline{\hat{\mathcal{P}}_{0:t}}(a), \tag{13}$$

where:

$$\underline{\hat{\mathcal{P}}_{0:t}}(a) \triangleq \min_{\hat{p}_{0:t}(X) \in \hat{\mathcal{P}}_{0:t}(X)} \hat{\mathcal{P}}_{0:t}(a) \tag{14}$$

$$\overline{\hat{\mathcal{P}}_{0:t}}(a) \triangleq \max_{\hat{p}_{0:t}(X)\in\hat{\mathcal{P}}_{0:t}(X)} \hat{\mathcal{P}}_{0:t}(a) \quad (15)$$

In this paper, we will use the *centroid,* i.e. the expected value with respect to a *second order uniform distribution* over $\hat{\mathcal{P}}_{0:t}(X)$, in order to perform the classification with respect to $T'$. Such decision schema can often be found in the literature discussing decision making based on credal sets [15].

# 4 Empirical Evaluation

To evaluate the performance and feasibility of the precise and imprecise anomaly detectors we design two experiments using real-world maritime AIS (Automatic Identification System) [18] data where deliberate anomalies are introduced.

## 4.1 Datasets

Laxhammar et al. [19] introduced a novel approach for comparing performance of anomaly detection methods, which we use in the first experiment. We use the same original dataset as in Laxhammar et al. [19] for all experiments. The dataset includes three weeks of AIS data recorded along the Swedish west coast. The dataset was preprocessed according to Section 3.3 in Laxhammar et al. [19] which resulted in a new dataset containing 2888 tracks with a total of 216717 observations. The dataset was divided into one training set (2310 tracks) and one evaluation set (578 tracks). Each track contains a number of observations with attributes such as; timestamp, latitude position, longitude position, speed, and heading. To suppress some of the complexity in the original dataset the observations were sampled with a distance of 200m [19]. The tracks are discretised into composite states $\mathbb{Y}$ consisting of atomic states: position, velocity and heading. We discretise by using a positional grid with 45 by 45 cells around the area of interest (each cell is 304 by 198 meters), four velocity classes with limits 3, 15 and 25 knots and eight equally large classes for the course. The parameters are set by a domain expert with respect to what can be expected in the domain. For more information regarding the discretization, see Brax et al. [7].

## 4.2 Experiment 1 - Detection Delay

The aim of this experiment is to evaluate the reactivity of the anomaly detectors i.e. *detection delay*. The experiment is equivalent to the experiment performed by Laxhammar et al. [19] where the authors compare two anomaly detection methods based on *Gaussian Mixture Model* (GMM) and *Kernel Density Estimator* (KDE). Assume that we for each track introduce anomalous observations after a random time point $t$ to the end of the track, by using a random walk function. We use the same dataset and random walk function as Laxhammar at al [19]. Since each modified track produces a constant flow of anomaly observations from $t$ to the end, we can classify the vessel as anomalous from that point. The question now is, how long it will take for our set of anomaly detectors to classify the vessel as anomalous from the given time point? Clearly, the more reactive anomaly detectors we construct the shorter delay for detecting the anomaly. However, by constructing highly reactive detectors we also increase the likelihood of classifying a normal vessel as anomalous before the vessel has started producing anomalous observations, resulting in a false alarm. As mentioned before the reactiveness of our anomaly detectors can be modeled by the parameters $s_a$, $s_n$ and $T$. In the experiments conducted by Laxhammar et al. [19], the threshold values for the GMM and KDE methods were set such that 1% of the normal trajectory segments contained one or more anomalous observations prior to time point $t$. This can also be interpreted as the classification of 1% of the evaluated tracks resulted in a false alarm, i.e. a normal tracks classified as anomalous. To be able to compare our proposed methods with the methods evaluated by Laxhammar et al. [19], we set the parameters $s_a$, $s_n$ and $T$ using a linear search, to generate at most 1% false alarms. We also set the parameters (see Section 2.1) of the sliding window method to generate at most 1% false alarms.

### 4.2.1 Results

The parameters that performed best for each method are shown in Table 1.

Table 1: Threshold settings for precise, imprecise and sliding window detectors. The parameter $n_{ext}$ corresponds to the number of extreme observations required to reach the threshold $T'$ starting from the prior evidence $\hat{p}_0(a)$ and $n_{ws}$ corresponds to the sliding windows size. The parameters for the sliding window method are described in Section 2.1.

| ***Detector*** | $\boldsymbol{n_{ext}}$ | $\boldsymbol{T}$ | $\boldsymbol{\hat{p}_0(a)}$ |
|---|---|---|---|
| Precise | 3 | $\min_{\mathbb{y}\in\Omega_{\mathbb{y}}} p(\mathbb{y})$ | 0.01 |
| Imprecise | 3 | $[\min_{\mathbb{y}\in\Omega_{\mathbb{y}}} p(\mathbb{y}), \min_{\mathbb{y}\in\Omega_{\mathbb{y}}} p(\mathbb{y})]$ | 0.01 |
| | | | |
| ***Detector*** | $\boldsymbol{n_{ws}}$ | $\boldsymbol{T}$ | $\boldsymbol{T_{al}}$ |
| Sliding Window | 5 | $\min_{\mathbb{y}\in\Omega_{\mathbb{y}}} p(\mathbb{y})$ | 3 |

Note that for this experiment we found that the optimal parameters were equal for the precise and imprecise detectors. The detection delay results for the methods using the parameters in Table 1 are shown in Table 2.

Table 2: Results from detection delay experiment. The table shows the mean and median number of observations after the time point $t$ that the detectors need to classify the track as anomalous. For the mean values a 95% confidence interval is also presented.

| *Detector* | *Mean* | *Median* | *False alarms* |
|---|---|---|---|
| Sliding window | $4.19 \pm 0.05$ | 3 | 0.94% |
| Precise/Imprecise | $3.60 \pm 0.05$ | 2 | 0.93% |

According to previous experiments [19], the GMM and KDE methods needed 17.72 and 17.43 observations respectively before classifying a track as anomalous. Both methods had a median of 12 observations. Our proposed precise and imprecise methods outperform the previous methods based on GMM and KDE and are slightly better than the sliding window method. A possible explanation of why the results of our proposed methods are better than the GMM and KDE is that our methods have the ability to capture the behavior over time, i.e. anomalies occurring in multiple time steps can be accumulated.

## 4.3 Experiment 2 – Precision and Recall

The aim of this experiment is to evaluate the precise and imprecise anomaly detectors ability to detect anomalies in the form of a sequence of anomalous observations. To evaluate the performance we employ two commonly used measures from *data mining* namely, *precision* and *recall* [2]. Assume that we know the true class of the tracks in the test dataset. We can then use such information in order to evaluate the performance of our detectors by using precision and recall:

$$Precision \triangleq \frac{n_a^t}{n_a^t + n_a^f} \tag{16}$$

$$Recall \triangleq \frac{n_a^t}{n_a^t + n_n^f} \tag{17}$$

where $n_a^t$ denotes the number of tracks that has been correctly classified as anomalous, $n_a^f$ denotes the number of tracks that has been misclassified as anomalous and $n_n^f$ is the number of tracks that has been misclassified as normal. Note that an optimal detector has a precision and recall of one.

The parameters for this experiment can be found in Table 3. Due to the possibility of anomalies in the training data we want to be able to adjust the sensitivity of the detectors in such a way that anomalies in the test data can be detected. In essence, a normal model $p(\mathbb{z})$ induces a partial ordering among the observations $\mathbb{y}$ with respect to normality. Such ordering can be used to determine the sensitivity of the detectors by setting the threshold $T$ (see Equation (6)) to $p(\mathbb{z})$ where $\mathbb{z}$ belongs to a specific level $T_{lev}$ in the ordering. Consider the following case:

$$\{\mathbb{z}_1^0, \dots, \mathbb{z}_m^0\} \prec \{\mathbb{z}_1^1, \dots, \mathbb{z}_{m'}^1\} \prec \{\mathbb{z}_1^2, \dots, \mathbb{z}_{m''}^2\} \prec \dots \tag{18}$$

where:

$$\mathbb{z}_\bullet^i \prec \mathbb{z}_\bullet^j \; iff \; p(\mathbb{z}_\bullet^i) < p(\mathbb{z}_\bullet^j) \tag{19}$$

If we set $T_{lev} = 2$, resulting in $T = p(\mathbb{z}_\bullet^2)$, then observations $\mathbb{z}_\bullet^0$ and $\mathbb{z}_\bullet^1$ yields anomaly evidences (see the mapping in Figure 1). Let us define three sets of parameter settings, denoted by *A*, *B*, and *C,* where we vary $T_{lev}$ for the precise detector. For the imprecise detector, we perform a search around each value of $T_{lev}$ by constructing the interval $[T_{lev} - \delta, T_{lev} + \delta]$ for a number of values of $\delta$ given in Table 3. Note that each such interval corresponds to $[\underline{T}, \overline{T}]$ (see Section 3.3).

We also vary $T_{lev}$ for the sliding window detector. Let us assume that ten consecutive observations that we have not seen before is regarded to be an anomaly. Hence, we set the parameter $n_{ext} = 10$ for both the precise and imprecise detectors. Similarly, for the sliding window detector we set the detection threshold $T_{al}$=10. The Window size parameter $n_{ws}$ is set to 20. This is mainly due to the results from the first experiments where the optimal window size was approximately $2 * T_{al}$. The lower evidence bound, $\hat{p}_0(n)$ is set to 0.01 for all methods.

Table 3: Threshold settings for the Precise, Imprecise and Sliding Window detectors.

| ***Detector*** | $\boldsymbol{n_{ext}}$ | ***A*** | ***B*** | ***C*** | $\boldsymbol{\hat{p}_0(a)}$ |
|---|---|---|---|---|---|
| Precise | 10 | $T_{lev} = 0$ | $T_{lev} = 5$ | $T_{lev} = 10$ | 0.01 |
| Imprecise | 10 | $\delta = 0$ | $\delta \in \{1,\dots,5\}$ | $\delta \in \{1,\dots,5\}$ | 0.01 |
| | | | | | |
| ***Detector*** | $\boldsymbol{n_{ws}}$ | ***A*** | ***B*** | ***C*** | $\boldsymbol{T_{al}}$ |
| Sliding Window | 20 | $T_{lev} = 0$ | $T_{lev} = 5$ | $T_{lev} = 10$ | 10 |

Now, in order to evaluate our choice of design of the detectors with respect to precision and recall, we need to construct a new test dataset where some tracks are anomalous. Since we have implicitly defined what we regard as an anomaly, i.e., ten consecutive observations not previously seen, let us introduce anomalies by skewing ten consecutive observations from some random start position. We perform the skewing of each point according to Figure 3. The maximum $\Delta$ is set to 500 m. Considering that the data is sampled every 200 m (see [19] for details), this will result in anomalies that are more subtle compared to the anomalies in the first experiment.

We use the same training dataset as in the first experiment and sample the evaluation dataset to construct the test dataset. The test dataset has 164 tracks labeled normal and 166 tracks labeled anomalous, i.e., tracks where we introduce anomalies as described above.

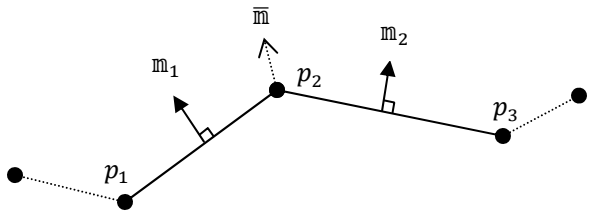


Figure 3: The point $\mathrm{p}_2$ becomes skewed to a new point $\mathrm{p}_2' = \mathrm{t}\,\bar{\mathbb{m}}$ where $\bar{\mathbb{m}} = (\mathbb{m}_1 + \mathbb{m}_2)/2$ and $\mathrm{t} \sim \mathrm{Un}([-\Delta, \Delta])$ ($\mathrm{Un}(\cdot)$ denotes the uniform distribution).

### 4.3.1 Results

The results are seen in Table 4. For the imprecise detector, the optimal intervals around $T_{lev}$ for the parameter settings B and C were found to be [1,9] and [5,15], respectively. We see that the precision of the imprecise detector is slightly better than the precise correspondence and that the performance with respect to recall is more or less the same. We also see that the sliding window method performs best with respect to precision (0.94). However, the recall for the corresponding setting A is low compared to setting B and C. In fact, parameter setting A is not a beneficial setting for any of the methods due to the low recall. We also see that as $T_{lev}$ increase, i.e. $T$ increase in Equation (6), the recall increase due to the fact that we obtain more anomalous evidences.

Table 4: The results from Experiment 2 for the different parameters settings *A*, *B* and *C*.

| ***Detector*** | ***Precision*** | | | ***Recall*** | | |
|---|---|---|---|---|---|---|
| | ***A*** | ***B*** | ***C*** | ***A*** | ***B*** | ***C*** |
| Precise | 0.91 | 0.88 | 0.85 | 0.30 | 0.98 | 0.98 |
| Imprecise | 0.91 | 0.92 | 0.88 | 0.30 | 0.97 | 0.98 |
| Sliding Window | 0.94 | 0.90 | 0.84 | 0.49 | 0.88 | 0.92 |

Even though the results do not differ that much between the precise and imprecise detectors, there exist tracks where the classification is clearly sensitive to the choice of $T$, resulting in a large intervals for the imprecise detector, see Figure 4.

## 5 Summary and Conclusions

We have extended the State-Based Anomaly Detection approach by introducing precise and imprecise anomaly detectors using the Bayesian and credal combination operators. We performed two experiments; the first one measured detection delay and the second one precision and recall. The results from the first experiment showed that our introduced detectors outperform previous detectors based on Gaussian mixture model and kernel density functions. Our proposed detectors performed slightly better than the sliding window detector. The results from the second experiment showed that the imprecise detector is slightly better than the precise and sliding window detector. The recall for all detectors was low without increasing the sensitivity using the threshold. This indicates a presence of anomalies in the training dataset.

One benefit with utilizing the precise and imprecise anomaly detectors, instead of the sliding window detector, is that it can be more intuitive to set the reactivity of a detector by setting the number of extreme observations needed for raising an alarm compared to setting a window size and a detection threshold for a number of binary detectors.

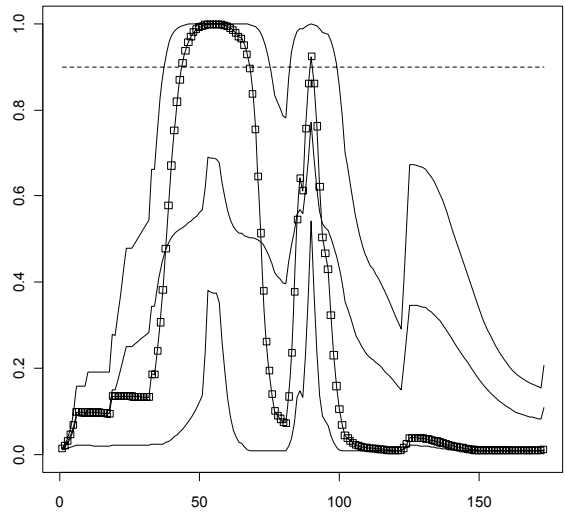


Figure 4: Example of the joint evidences $\hat{p}_{0:t}(a)$ (squares) and $\hat{\mathcal{P}}_{0:t}(a)$ for the precise and imprecise detectors based on $p_{Occ}(\mathrm{y})$ (see Equation 1). The centroid of $\hat{\mathcal{P}}_{0:t}(a)$ is also shown. The y-axis depicts the strength of the evidence and the x-axis show the time step $t$.

One potential advantage of utilizing the imprecise detector, instead of the precise, is that one can use imprecision as a way of conveying information about the reliability of an classification to a human operator. In essence, when an operator determines a set of thresholds, the operator also agrees on using a set of detectors for performing the classification. If all these detectors classify the object as an anomaly then the operator can rely on such results more than if there only is a subset of detectors that do so. One way this can be utilized is to give an early warning to the operator when the detector corresponding to the upper threshold reaches the classification threshold. The human operator can in such cases inspect the amount of imprecision in order to obtain information about the reliability.

For future work we want to evaluate the proposed detectors on another dataset with more features and where

the area of interest is more cluttered with objects (cf [8]). For such dataset, it can perhaps be more beneficial to utilize the imprecise detector, in particular when there is a human operator involved in the decision process.

## Acknowledgements

This work was supported by the Information Fusion Research Program (University of Skövde, Sweden) in partnership with the Swedish Knowledge Foundation under grant 2003/0104 (URL: http://www.infofusion.se). This work was also supported by Saab AB, Sweden.

## References

[1] National Plan to Achieve Maritime Domain Awareness, Department of Homeland Security, USA, 2005.

[2] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, Addison-Wesley, Boston, 2006.

[3] L. Portnoy, E. Eskin, and S.J. Stolfo, Intrusion detection with unlabeled data using clustering, ACM Workshop on Data Mining Applied to Security (DMSA-2001), USA, 2001.

[4] S. Arnborg, Robust Bayesianism: Relation to Evidence Theory. Journal of Advances in Information Fusion 1 (2006) 63-74.

[5] S. Arnborg, Robust Bayesianism: Imprecise and Paradoxical Reasoning, Proceedings of the 7th International Conference on Information Fusion, Stockholm, Sweden, 2004.

[6] A. Karlsson, R. Johansson, and S.F. Andler, On the Behavior of the Robust Bayesian Combination Operator and the Significance of Discounting, 6th International Symposium on Imprecise Probability: Theories and Applications, Durham, U.K., 2009.

[7] C. Brax, and L. Niklasson, Enhanced Situational Awareness in the Maritime Domain: An Agentbased Approach for Situation Management, SPIE Security, Defence + Sensing 2009, Orlando, Florida, USA, 2009.

[8] C. Brax, L. Niklasson, and R. Laxhammar, An ensemble approach for increased anomaly detection performance in video surveillance data, The 12th International Conference on Information Fusion, Seattle, WA, USA, 2009, pp. 694-701.

[9] C. Brax, and L. Niklasson, An approach for increased supply chain security by using automatic detection of anomalous vehicle behaviour, Modelling Decisions for Artificial Intelligence, Awaji Island, Japan, 2009.

[10] C. Brax, R. Laxhammar, and L. Niklasson, Approaches for Detecting Behavioural Anomalies in Public Areas Using Video Surveillance Data, SPIE Europe Security + Defence 2008, Cardiff,.Wales, United Kingdom 2008.

[11] C. Brax, L. Niklasson, and M. Smedberg, Finding behavioural anomalies in public areas using video surveillance data, The 11th International Conference on Information Fusion, Cologne, Germany, 2008.

[12] J.e. Bernardo, M., and A.F.M. Smith, Bayesian Theory, John Wiley and Sons, 2000.

[13] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, 1991.

[14] I. Levi, The enterprise of knowledge, The MIT press, 1983.

[15] F. Cozman, Decision Making Based on Convex Sets of Probability Distributions: Quasi-Bayesian Networks and Outdoor Visual Position Estimation, The Robotics Institute, Carnegie Mellon University, 1997.

[16] F.G. Cozman, Graphical models for imprecise probabilities. International Journal of Approximate Reasoning 39 (2005) 167-184.

[17] I.e. Couso, s, S.i. Moral, n, and P. Walley, A survey of concepts of independence for imprecise probabilities. Risk Decision and Policy 5 (2000) 165-181.

[18] T. Eriksen, G. Hoye, B. Narheim, and B.J. Meland, Maritime traffic monitoring using a space-based AIS receiver. Acta Astronautica 58 (2006) 537-549.

[19] R. Laxhammar, G. Falkman, and E. Sviestins, Anomaly detection in sea traffic - a comparison of the Gaussian Mixture Model and the Kernel Density Estimator, The 12th International Conference on Information Fusion, Seattle, WA, USA, 2009, pp. 756-763.