Online Maritime Abnormality Detection using Gaussian Processes and Extreme Value Theory

Mark Smith
ISSG, Babcock Marine & Technology Division,
Devonport Royal Dockyard,
Plymouth, United Kingdom
mark.x.smith@babcockinternational.com

Steven Reece, Stephen Roberts, Iead Rezek
Department of Engineering Science,
University of Oxford,
Oxford, United Kingdom
{*reece, sjrob, irezek*}*@robots.ox.ac.uk*

***Abstract*—Novelty, or abnormality, detection aims to identify patterns within data streams that do not conform to expected behaviour. This paper introduces a novelty detection technique using a combination of Gaussian Processes and extreme value theory to identify anomalous behaviour in streaming data. The proposed combination of continuous and count stochastic processes is a principled approach towards dynamic extreme value modelling that accounts for the dynamics in the time series, the streaming nature of its observation as well as its sampling process. The approach is tested on both synthetic and real data, showing itself to be effective in our primary application of maritime vessel track analysis.**

***Keywords*-Gaussian Process; Extreme Value; Maritime Traffic; Novelty Detection; Outlier Detection.**

I. Introduction

The global picture of maritime traffic is large and complex, consisting of dense volumes of (mostly legal) ship traffic. Techniques that identify illegal traffic could help to reduce the impact from smuggling, terrorism, illegal fishing etc. In the past, surveillance of such traffic has suffered due to a lack of data. However since the advent of electronic tracking the amount of available data has grown beyond an analyst's ability to process without some form of automation. One part of the analyst's workload lies in the detection of anomalous behaviour in otherwise normal appearing tracks. Our goal is to detect anomalous vessels using an automated approach. In this paper we exploit techniques from the field of anomaly detection, particularly *extreme value theory* to identify potential deviations from normal behaviour. The latter is modelled using a non-parametric Bayesian approach, namely sequential *Gaussian Process* regression.

An anomaly has many different interpretations depending on the context in which it is used, i.e. it may refer to a data point arising from: a different distribution, measurement error, population variability or execution error. However, fundamentally an anomaly is a data point which stands out in contrast to the other data points around it [1]. It is the task of anomaly detection to infer whether this data point deviates significantly given the intrinsic variability of the population of normal data. An effective technique should therefore be capable of recognising and modelling data points that occur due to such anomalous events and distinguish these from outliers associated with the tails of the reference distribution of non-anomalous data.

This paper applies *extreme value statistics* to identify likely anomalous samples which are extreme values. The probability distribution governing these extreme values is sequentially updated to enable context sensitive decisions. This is achieved by means of linking this distribution to a sequential Gaussian Process model which regresses the vessel's track and forecasts a distribution over future data. This paper begins by providing a brief overview of existing approaches to maritime situational awareness, followed by a description of our methods. An application to synthetic data is included to provide insights into the workings of the approach. An application to real vessel tracks extracted from their GPS co-ordinates is then presented.

II. Current Techniques

The field of marine anomaly detection has employed a variety of methods including Neural Networks [2], Bayesian Networks [3], Support Vector Machines [4], GPs [5] and Kalman filters [6]. Common between all methods are two main tasks; creating a model of normality (free from the presence of anomalies), and using a metric from this model which (allowing for some quantifiable variability) identifies a point as an anomaly. These tasks are inherent within the two main uses for the detection of anomalies; *accommodation* and *discordancy*. Accommodation is the task in which the goal is create a model of normality that does not include anomalous observations. Whereas discordancy tests provide a metric indicator of a point being an anomaly. Both have different aims but within each is some model of normality, and a measure of deviation.

Assessing the performance of these different methods is a difficult task as there exist no established benchmarks of what are considered as marine anomalies, therefore hindering comparison [7]. This implies that a data set can be considered under a variety of contexts, leading to different types of anomalies being identified on the same input data.

For example a time series analysis of a vessel track, where the previous location and the dynamics of the vessel are considered, could highlight sudden changes in the vessel dynamics; a sudden change possibly indicative of evasive manoeuvring. Such a time series model has the advantage that it can be used in online analysis, but it may miss patterns when the data set is considered simultaneously [3]. Other indications of anomalous behaviour within the data could be deviations from standard route, unexpected port arrival, close approach and zone entry [8]. Even when a particular marine anomaly has been selected for identification, the data needs to be considered in the context of external variables, for example the class of vessel, time of day, tidal status. Since these may have to be taken into account when analysing the data as the form of the anomaly may vary.

Critical to marine anomaly detection is an interpretation of the data that allows the salient features of the desired anomaly to be identified [9]. Models for different kinds of anomalies may need to be combined or considered to increase the certainty of an anomaly being detected. For example a model identifying anomalous vessel speeds could be combined with a model of anomalous zone entries (anomalous spatial locations). Vessels identified as demonstrating anomalous speeds may be pleasure craft in a known unrestricted speed location, giving false positives if simply considered only on the basis of the speed model. Conversely a vessel entering a port at high speed may be highly anomalous behaviour.

III. The Gaussian Process-Extreme Value (GP-EVT) Approach

The approach we detail in this paper develops methods for the two main underlying tasks within anomaly detection; modelling normality and subsequently using a cost function to identify points as anomalous. We construct a model of normality using Gaussian Processes (GPs) which allows us to capture the dynamics of vessels in a non-anomaly data set without prescribing a particular parametric form (such as is required for Markov state models, for example). The GP provides a sequentially updated posterior distribution over unseen data, which we link to an extreme value distribution to provide a robust and adaptive metric for anomaly detection.

A. Gaussian Processes

In order to model the vessel track we use a Gaussian Process. It provides a mechanism which we use to continuously predict vessel locations at any future time point, *including* a measure of uncertainty about the vessel location. The GP is a stochastic process [10] that expresses the dependent variable, y, in terms of an independent variable x, via a function $f(x)$. This function we can see as a draw from a probability distribution over functions,

$$y = f(x) \sim \text{GP}(m(x), k(x, x)), \tag{1}$$

where $m(x)$ describes the mean function of the distribution and k is a covariance function which describes the information coupling between two values of the independent variable as a function of the distance of their respective inputs. This covariance function thus encodes our beliefs and assumptions about the function that we wish to model [10]. Valid covariance functions can take a variety of forms which we quantify empirically in this paper. Denoting $r = |x_p - x_q|$ as the (Euclidean) distance between two independent variable points, x_p and x_q, we consider three covariance functions: the squared exponential

$$k_{SE}(r) = \sigma_0^2 \text{exp}\left(-\frac{r^2}{2\lambda^2}\right); \tag{2}$$

the Matérn $\frac{3}{2}$

$$k_{\frac{3}{2}}(r) = \sigma_0^2 \left(1 + \frac{\sqrt{3r}}{\lambda}\right) \text{exp}\left(-\frac{\sqrt{3r}}{\lambda}\right); \tag{3}$$

and the Matérn $\frac{1}{2}$ covariance function

$$k_{\frac{1}{2}}(r) = \sigma_0^2 \text{exp}\left(-\frac{r}{\lambda}\right). \tag{4}$$

The above selection was driven by prior knowledge about typical vessel trajectories. For example, periodic kernels [10] were excluded from the set of covariance functions as the data in our feature space is not periodic. Hence, our choice was restricted to covariance functions capable of reflecting the physical properties of shipping vessels, such as smoothness and differentiability.

We also assume that the observations are corrupted by additive i.i.d Gaussian noise with variance component ϵ^2. Thus, the full covariance function is given as

$$V(x_p, x_q) = k(x_p, x_q) + \epsilon^2 \delta(|x_p - x_q|), \tag{5}$$

where δ is the Kronecker delta, which is one if $p = q$ and zero otherwise.

The hyperparameters σ_0, λ and ϵ are referred to, respectively, as the amplitude, output and noise scale. They encode the characteristics of the track and so depend on the dynamics of the vessel. A vessel undertaking manoeuvring will not exhibit the same smooth track characteristics as one exhibiting regular motion. Thus, the hyperparameters need to be learnt from an anomaly free training data set which consist of n observations, $D = \{(x_i, y_i)|i = 1, ..., n\}$. The x_i and y_i points represent the independent and dependent variable values respectively.

The nature of the Gaussian Process is such that, conditional on observed data, predictions can be made about the function values, $f(x_*)$ at any location x_*. The distribution

of these values at point x_* is Gaussian with mean and covariance, given as

$$f_*|x_*, x, y \sim \mathrm{N}(\bar{f}_*, \mathrm{Var}[f_*]). \tag{6}$$

This gives rise to the following predictive equations for GP regression, for which we assume the mean function m to be zero,

$$\begin{aligned} \bar{f}_* =& m(x_*) + k(x_*, x_*)^\top V(x,x)^{-1}(y - m(x)), \\ \mathrm{Var}[f_*] =& k(x_*, x_*) - k(x, x_*)^\top V(x,x)^{-1} k(x, x_*). \end{aligned} \tag{7}$$

B. Sequential Gaussian Process Updates

In real world problems we receive data sequentially and the total data set can grow to arbitrarily large size. If we were to continue to update our beliefs in the light of new observations we could naively repeat the matrix inversion in Equation 7 with every observation. This inversion is expensive as its computational complexity grows as $O(n^3)$ in the number of samples, i.e. the dimension of the matrix, V above. Closer inspection however, reveals that covariance matrix V is changed only in the addition of some new rows and columns. Hence, it is possible to reformulate the matrix inversion as a sequential Cholesky decomposition [11].

We decompose a matrix into the product of a lower triangular matrix, R, and its conjugate transpose

$$V(x,x) \triangleq R(x,x)^\top R(x,x). \tag{8}$$

Based on this decomposition, the predictive distribution is given as

$$\begin{aligned} \bar{f}_* =& m(x_*) + b_{x,x_*}^\top a_x V(x_*, x_*), \\ \mathrm{Var}[f_*] =& V(x_*, x_*) - b_{x,x_*}^\top b_{x,x_*}, \end{aligned} \tag{9}$$

where a and b are given as

$$\begin{aligned} a_x \triangleq& R(x,x)^\top \backslash (y - m(x)), \\ b_{x,x_*} \triangleq& R(x,x)^\top \backslash V(x, x_*). \end{aligned} \tag{10}$$

When we receive new data, the V matrix is changed only in the addition of some new rows and columns, i.e.

$$V(x,x) = \begin{pmatrix} V(x_{1:n-1}, x_{1:n-1}) & V(x_{1:n-1}, x_n) \\ V(x_n, x_{1:n-1}) & V(x_n, x_n) \end{pmatrix}. \tag{11}$$

Consequently the Cholesky decomposition can also be computed iteratively [11] as

$$R(x_{1:n}, x_{1:n}) = \begin{pmatrix} R(x_{1:n-1}, x_{1:n-1}) & S \\ 0 & U \end{pmatrix}, \tag{12}$$

where

$$\begin{aligned} S =& R(x_{1:n-1}, x_{1:n-1})^\top \backslash V(x_{1:n-1}, x_{1:n}), \\ U =& \mathrm{chol}(V(x_n, x_n) - S^\top S). \end{aligned} \tag{13}$$

With this Cholesky update expressed iteratively the predictive distribution, Equation 9, can also be expressed iteratively by expressing the vector a, in Equation 10, via the simple update rule

$$a_{1:n} = \begin{pmatrix} a_{1:n-1} \\ U^\top \backslash (y_n - m(x_n) - S^\top a_{1:n-1}) \end{pmatrix}. \tag{14}$$

This avoids the computationally expensive matrix inversion, in Equation 7, and allows the Cholesky factor to be expressed as an efficient update rule. We now consider a principled means of determining whether a new data point should be updated into the model of normal system behaviour.

C. Extreme Value Theory

Extreme value theory has previously been used to create a novelty detection threshold, [12], [13], beyond which we can quantify a value as having not arisen from the underlying distribution. The theory itself focuses on the statistical behaviour of $M_n = \max\{X_1, \ldots, X_n\}$ where $X_1, \ldots, X_n$ is a sequence of independent random variables with a distribution function F. In theory the distribution of M_n can be derived exactly for all values of n, i.e.

$$\begin{aligned} Pr\{M_n \leq z\} =& Pr\{X_1 \leq z, \ldots, X_n \leq z\} \\ =& Pr\{X_1 \leq z\} \times \ldots \times Pr\{X_n \leq z\} \\ =& \{F(z)\}^n. \end{aligned} \tag{15}$$

In practice the distribution function F is unknown and extreme value theory allows us to approximate this distribution. It states that the entire range of possible limit distributions for M_n is given by one of three types of cumulative distribution function, I, II and III, known as the Gumbel, Fréchet and Weibull, respectively, and given as

$$I : G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\} \quad -\infty < z < \infty \tag{16}$$

$$II : G(z) = \begin{cases} 0, & z \leq b, \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b \end{cases} \tag{17}$$

$$III : G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\alpha}\right]\right\}, & z < b \\ 1, & z \geq b \end{cases} \tag{18}$$

Each family has a scale and location parameter, a and b respectively. Additionally the Fréchet and Weibull families have a shape parameter α [14]. Although we have three models to choose from, the underlying target distribution, F, in our case is assumed to be Gaussian, due to the modelling constraints imposed by the GP. The extreme value probability is then restricted to the analytical form of the Gumbel distribution.

Assuming that some "normal" data is identically and independently Gaussian distributed, one can obtain the extreme quantiles by inverting Equation 16

$$z_p = b - a \log\left(-\log\left(p\right)\right). \tag{19}$$

The value of p acts as a novelty threshold, below which a test point is classified "abnormal". The parameters a and b require estimation and typically depend on the sample size n of the data set. As proposed in [12], we make use of decoupled estimators for a and b given, respectively as

$$a = \left(2\log\left(n\right)\right)^{-\frac{1}{2}} \tag{20}$$

$$b = \left(2\log\left(n\right)\right)^{\frac{1}{2}} - \frac{\log\left(\log\left(n\right) + \log\left(2\pi\right)\right)}{2\left(2\log\left(n\right)\right)^{\frac{1}{2}}} \tag{21}$$

D. Gaussian Process Extreme Value Theory

In much of the existing work on novelty detection using extreme value methods, the work has focused on non-sequential conditions, or more precisely, on a fixed training dataset. While the extreme value will adequately account for the changes in our belief about the location of extreme events for a fixed sample size, the framework is rarely extended to account for dynamic changes in the underlying generating distribution *and* changes in the sample size.

In this work we model the typical system dynamics using Gaussian Process (GP) regression. At some arbitrary point in the future, say x_*, we can interrogate the GP and compute the predictive (Gaussian) distribution at that point, conditional on the trajectory's past samples. This predictive distribution, which now features a context (time) dependent mean, $\bar{f}_*$, and variance, $\mathrm{Var}[f_*]$, allows rescaling of the extreme event quantile e,

$$e = \bar{f}_* + \sqrt{\mathrm{Var}[f_*]} z_p. \tag{22}$$

and so reflect temporal changes in the statistics of the base distribution.

In order to estimate the number of data points $n(x_*)$ at each x_* in Equation 9, a Gaussian kernel smoother is applied to compute the predicted number of points at x_*, using

$$n(x_*) = \sum_{i}^{n} \phi_h\left(x_*; x_i\right), \tag{23}$$

where $\phi_h(x_*; x_i)$ is a (non-normalised Gaussian) radial basis function

$$\phi_h\left(x_*; x_i\right) = \exp\left\{-\frac{|\, x_* - x_i\,|^2}{2h^2}\right\},$$

x_i is the most recent observation and $|\cdot|$ denotes the Euclidean distance. The kernel width h is set to be equal to twice the length scale λ in Equations 2, 3 and 4, depending on the choice of kernels used to model the vessel tracks. This coupling of λ to the GP regression models ensures that tracks with long correlation lengths and smaller sampling rates will feature the same sensitivity to outliers as tracks with short correlation lengths and high sampling rates. Also, the coupling ensures that the smoothing of the sampling processes does not come at a cost of an additional parameter which would require additional estimation or ad hoc choice.

With the expected number of observations obtained by Equation 23, the extreme value distribution parameters can be updated in a timely fashion to reflect also the dynamics of the sampling process. Thus, the scaling (Equation 20), and location (Equation 21), parameters can be estimated, using the predicted number of data points, n, contributing information at the location of interest x_* [15], by

$$a = \left(2\log\left(n(x_*)\right)\right)^{-\frac{1}{2}} \tag{24}$$

and

$$b = \left(2\log\left(n(x_*)\right)\right)^{\frac{1}{2}} - \frac{\log\left(\log n(x_*)\right) + \log\left(2\pi\right)}{2\left(2\log\left(n(x_*)\right)\right)^{\frac{1}{2}}} \tag{25}$$

for a fixed novelty detection threshold, p, which in this work is set to 0.95.

To reiterate, the GP provides a mechanism to predict the distribution of future mean values and to adjust the scaling of the extreme value quantile. Also, the kernel smoothing approach to the sampling process provides an estimate of the future sample size. Their combination is used for novelty detection. If the new data point value falls within a novelty measure of the predicted value then the new data point is included in the model update. The key advantage of using our approach is thus the incorporation of future uncertainty in both sampling and observation processes to provide the means for a more accurate novelty detection algorithm. The graphical model representation of the complete model is shown in Figure 1.

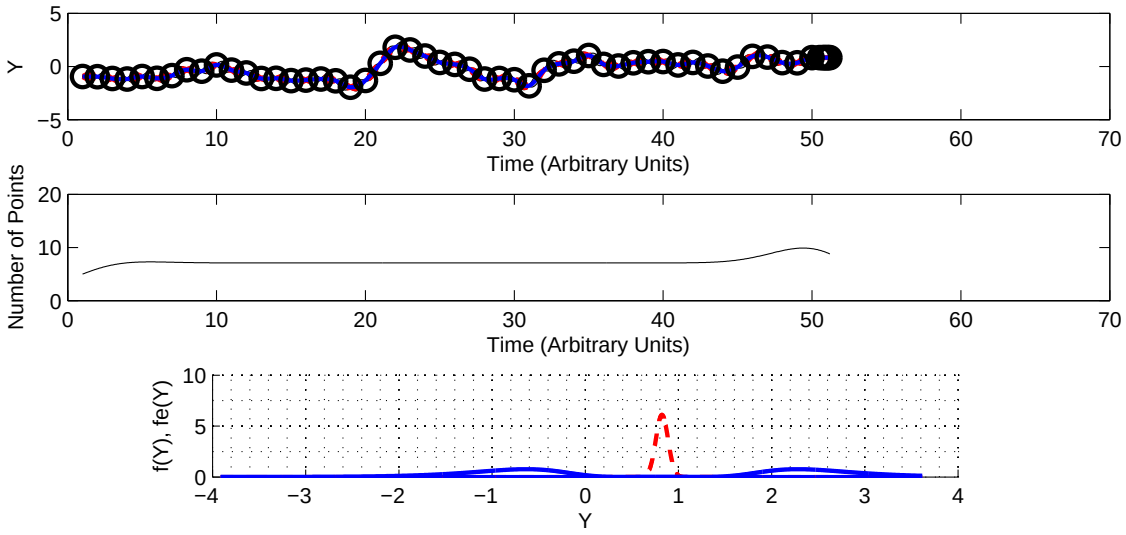
IV. Application

We demonstrate the efficacy of the approach presented in the previous section by application to synthetic data and real data. We use synthetic data to illustrate some of the features of our method, and provide a real world example of its application to vessel tracks.

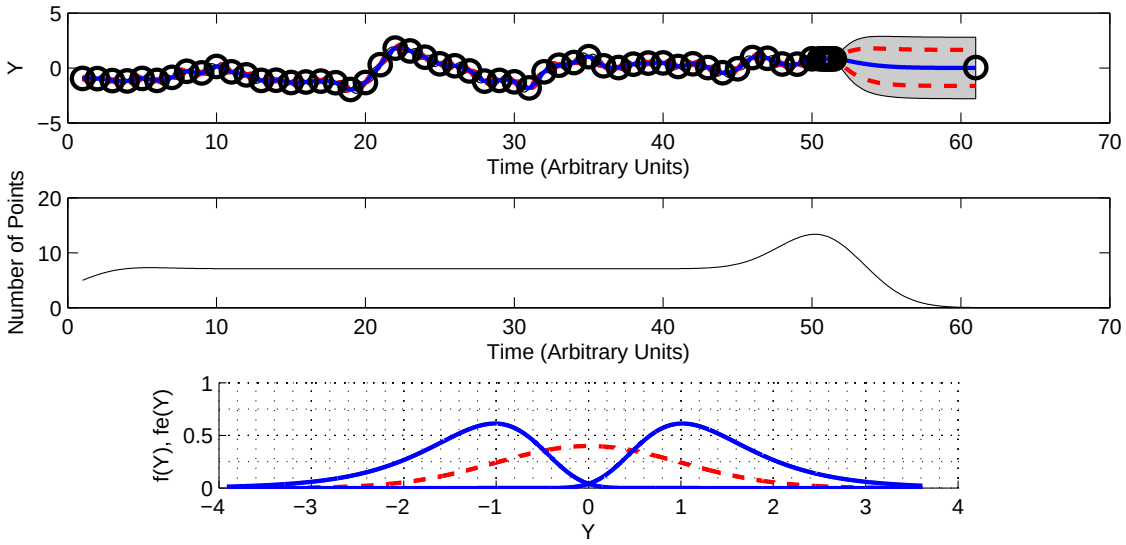
A. Synthetic Data Illustration

Synthetic data was generated from the Matérn$\frac{3}{2}$ kernel, Equation 3, with parameters set to $\sigma_0 = 1$, $\lambda = 2$ and $\sigma = 0.01$. Anomalies were generated by offsetting randomly selected samples that were previously drawn from the GP by a fixed offset value, making the point anomalous with respect to surrounding points. The GP predictive distribution was calculated for 1000 samples within the windowed region of track, the window ending at the time period for the new observed sample. A fixed kernel width was used in order to estimate n at each x_*.

GP extreme value theory was then applied, each new data point was considered with respect to the previously learnt underlying function. If the new point falls within the



(a) The sequential GP-EVT stopped at a region of high observation density. The estimate of the number of data points which contribute to the GP inference has increased significantly, as indicated by the observation density shown in the middle plot. Consequently the location of the extreme value distributions, illustrated by the continuous lines in the bottom plot, move away from the posterior predictive distributions (dashed line).



(b) The sequential GP-EVT (continuous line, upper plot), stopped at a region of very low observation density. The predictive distribution, (dashed line) is an accurate representation of the true distribution (continuous line). This is due to the relationship between the observation density and the location and scaling of the extreme value distribution, expressed in Equations 24 and 25.

Figure 2: Simulation of the effect of varying observation density on the extreme value distribution and hence the anomaly detection. Both plots show a snapshot from the last observed sample. In 2a the observation rate is high, while in 2b the observation rate is low. The observation density affects the location of the probability density function of the extreme value distribution $f_e(y)$ (blue lines, lower plot), relative to the predicted Gaussian PDF $f(y)$ (red dashed line, lower plot), drifting closer to the base distribution as the number density of points decreases.

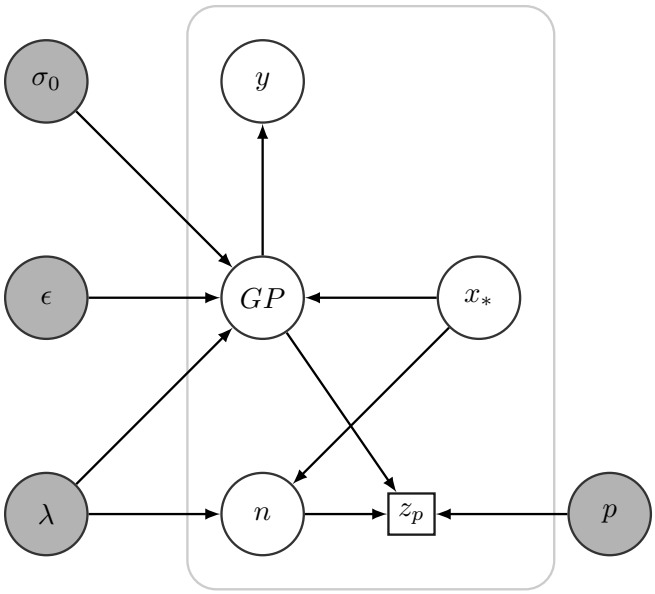


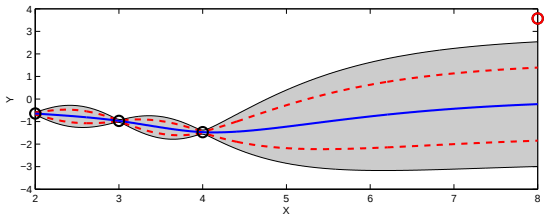
Figure 1: A graphical model representation of the GP-EVT model. At the centre is the Gaussian Process (GP) which models the track's dynamics. It has fixed hyper-parameters shown as grey nodes to the left. Also shown is the estimate of the sample size, n. The extreme value percentile is a deterministic node, shown as a square box, which depends upon p, the novelty level, and sample size n.

predictive uncertainty of the next data point it is included in the sequential update otherwise it will be excluded. An example of such an update step is shown in Figure 3. The new data point falls outside the EVT bound, Equation 22, and so has been excluded. Notice, that if a data point has not been observed for a period, the predictive uncertainty grows, allowing for the possibility of a dynamic change in the underlying base function and the new data point to be included in the update. In this manner anomalous points within the data can be clearly identified while perfectly accommodating for the dynamics of the underlying function and the irregular nature of its observation. An intuitive illustration of how the irregularity of observed data points effects the scaling of the extreme value distribution, and hence our novelty bounds is given in Figure 2.

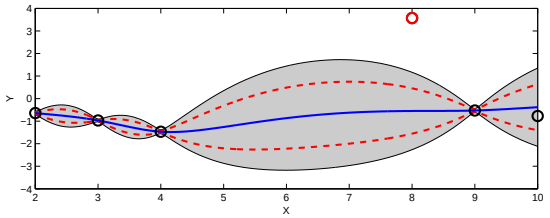
B. Vessel Track Anomaly Detection

The methodology was also applied to real world vessel track data. The data consists of a set of GPS coordinates which were collected from *marinetraffic.com*. In order to detect changes in the ship dynamics the data had to be converted to a appropriate feature space that provides a means of identifying such deviations. This feature space we describe next.

Feature Extraction: In order to convert data to a sufficient feature space representation we consider the first



(a) The GP predicts forward to the new artificially perturbed data point, and by using GP-EVT the new observation is classified as an anomaly.



(b) GP predicts forward after detecting and excluding the artificial anomaly. The uncertainty bounds continue to increase (until they reach their maximum as set by the prior distribution). The subsequent observations fall well within the error bound and so will be included in the next update.

Figure 3: Simulation of the GP prediction and anomaly detection. The continuous line shows the predicted mean function and the grey areas show the EVT bound of the GP predictive distribution for $p = 0.95$. The bound is open to the right and widening until the next observation has been included. Once it has, the standard deviation bound of the GP is updated and results in the familiar pointed elliptical shape, as seen between time steps 4 and 9. The dashed line shows the error bound produced if we consider the 95% bound from the mean function (1.64 standard deviations from the mean).

received data point as the beginning of the vessel track. We relate all subsequent data points to it by computing both the distance and time taken from this originating sample point. In order to take into account the approximated spherical geometry of the earth's surface we calculate this distance by application of the Haversine formula

$$\begin{aligned} A &= \cos\phi_s \cos\phi_f \\ \Delta\hat{\sigma} &= \arctan\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + A\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \end{aligned} \tag{26}$$

where ϕ_s and ϕ_f are the latitude of two points and $\Delta\lambda$ and $\Delta\phi$ are their differences in longitude and latitude respectively. This choice of feature space has the advantage of converting the GPS information into a 1D feature vector, reducing the computational demands of processing the data. Also, the arc length between points d for a sphere of radius r and $\Delta\hat{\sigma}$ given in radians by

$$d = r\Delta\hat{\sigma}. \tag{27}$$

Choice of Covariance Functions: The choice of covariance function is crucial in the methods ability to provide the most accurate representation of the vessel dynamics. To determine the optimal covariance function we investigated the performance of the the standard squared exponential kernel, in Equation 2, Matérn$\frac{3}{2}$, in Equation 3, and the Matérn$\frac{1}{2}$ kernel, Equation 4. Clean, i.e. anomaly free training data, was extracted from the training corpus and the GP kernel function parameters were estimated by maximising the marginal likelihood of the data [10]. The so obtained scores were standardised to the training data length and are shown in Table I.

Matérn$\frac{3}{2}$	Matérn$\frac{1}{2}$	SE
0.3348	0.3323	0.3334

Table I: Table of likelihood scores for the Matérn$\frac{3}{2}$, Matérn$\frac{1}{2}$ and standard squared exponential kernels.

The results suggest almost comparable performance, in terms of goodness of fit, of all three tested covariance functions. However, as shown in Figure 4, there is a substantial difference in the robustness. The Matérn$\frac{1}{2}$ kernel frequently finds poorer fits to the data. The squared exponential performs in the middle range, occasionally finding worse solutions than the Matérn$\frac{3}{2}$ kernel but better than the Matérn$\frac{1}{2}$ kernel.

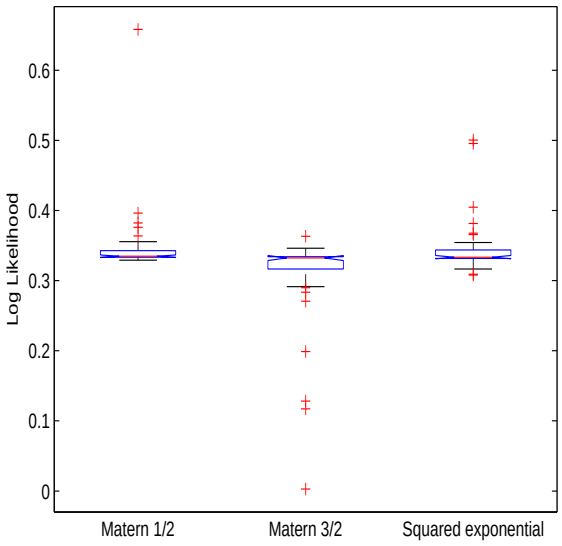


Figure 4: Log scores for the different covariance functions applied to each track.

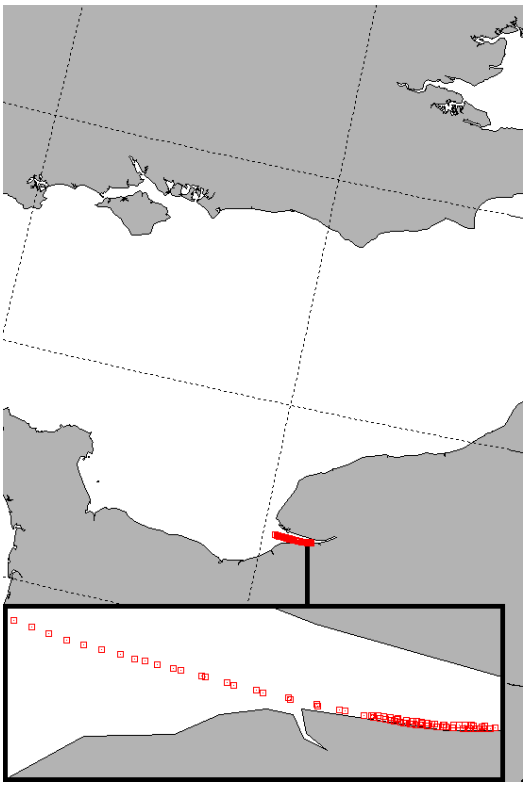
Vessel Track Modelling: The methodology was also applied to real world vessel track data. The Matérn $\frac{3}{2}$ kernel was chosen to model the underlying dynamics using hyperparameters learnt from anomaly free training data, which were chosen to be sufficiently long enough so that the underlying dynamics of the vessels could be captured.

Figure 5 shows an example vessel track without outlying points. The track is from a dredger which follows a smooth trajectory and does not make any sudden changes in acceleration. Shown in Figure 5b are the sequential EVT bounds sea-sawing until the next observation arrives. All observations fall well inside the predictive boundary of the GP-EVT bound and, consequently, no anomalies are detected.

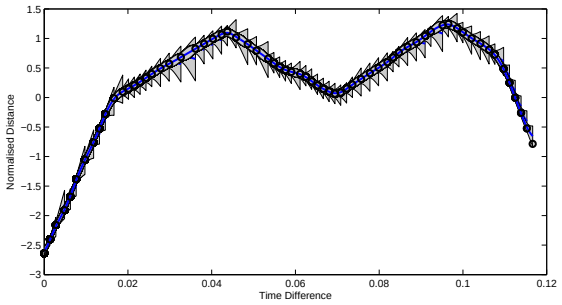
Figure 6 shows an example of a vessel track with some points which our model labels as anomalies. As can be seen in Figure 6a, the vessel remains within a confined area and there are short sudden movements, Figure 6b. These are marked as anomalies and are perhaps the result of the vessel drifting, manoeuvring or being moored. Figure 6c also shows an enlarged section of the sequential computation of the EVT bound and makes clear the non-linear relationship between the GP standard deviation bound and the actually computed EVT bound which includes essential parameters such as the number of predicted observations.

V. Comparison of GP-EVT with a Traditional Kalman Filter Approach

In this section we compare a traditional approach to anomaly detection with our Gaussian process and EVT approach. The traditional approach uses a Kalman filter (KF) to model the normal behaviour of the ship and then determines that the data is anomalous if it is more than a fixed number of standard deviations from the mean [16] (typically 3 to 5 standard deviations). This approach to anomaly detection is called the *gating* approach. Further, the KF approach requires a process model of the normal behaviour of the ship. Typically, a near constant velocity model is chosen to model the continuous trajectory without imposing any excessive smoothness on the trajectory [17]. We compare our outlined approach to a KF which uses the constant velocity model. This provides a fair comparison as both Matérn$\frac{3}{2}$ and constant velocity model are second order differentiable. We further investigate both a traditional KF using the standard deviation gating approach to exclude anomalies and a KF which uses the EVT in a manner similar to the GP. In so doing, we are able to compare both models of normal ship behaviour (namely the Matérn$\frac{3}{2}$ and the near constant velocity model) and also both approaches to detecting and excluding anomalies (namely, the EVT and standard deviation gating approaches). When using the KF the mean and standard deviation were predicted forward to the same time step as the new observation. If the point lies within a pre-chosen confidence region (defined as a multiple of the standard deviation about the mean) it is included in the update. This was repeated for a range of confidence regions defined by different multiples of the



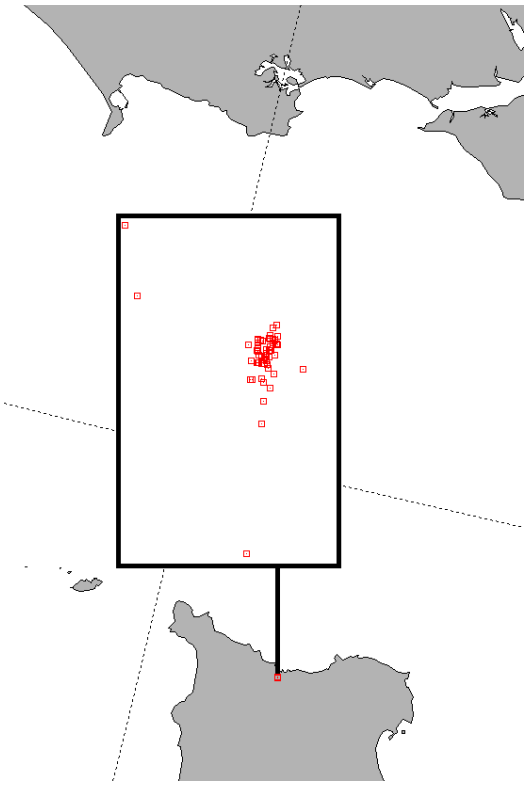
(a) A plot of the GPS track in which there were no detected anomalies.



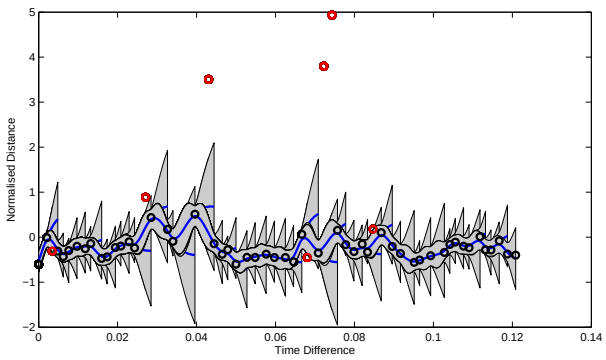
(b) Sequential predictions applied to feature extracted data, also showing that all data points fall within the EVT bound.

Figure 5: Sequential GP-EVT method applied to a dredging vessel operating off the coast of France near Le Havre.

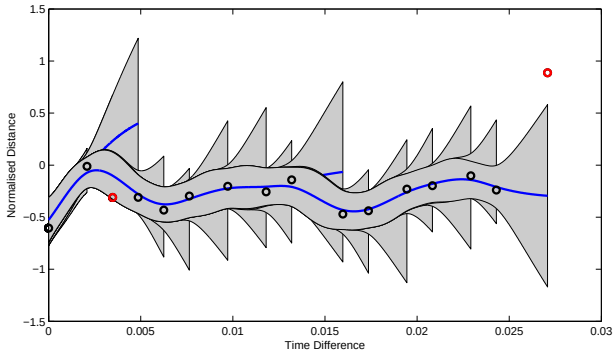
standard deviation. ROC curves were plotted for the results. The resulting area under curve which compares the KF using the near constant velocity model (with and without EVT) against the GP using a Matérn$\frac{3}{2}$ model (again with and without the EVT) are shown in Table II. We note that both the KF and GP performance is significantly improved using the EVT as opposed to gating. This is due to the fact that the gating approach uses a fixed threshold which does not take into account the density of observations i.e as we observe more samples we gain a better understanding of



(a) A plot of the GPS track in which there were several detected anomalies.



(b) Sequential predictions applied to feature extracted data, also showing some data points that fall outside the GP-EVT bound.



(c) Magnified section of the plot in Figure 6b. The GP-EVT bound has predicted forward to the new data point and included the point in the update.

Figure 6: Sequential GP-EVT method applied to a small vessel operating off the coast of France near Cherbourg and whose track suggests unusual navigation behaviour.

the true distribution of values. The EVT, however, uses a dynamic threshold which takes into account the density of observations therefore better utilises available information to adjust the threshold.

GP-EVT	GP	KF-EVT	KF
0.8032	0.7889	0.6545	0.6119

Table II: AUC for KF using the near constant velocity model (with and without EVT) and GP using a Matérn$\frac{3}{2}$ model (again with and without the EVT).

Although the results indicate a significant improvement of our model over the KF approach this is a limitation of the near constant velocity model used and not a critique of KF based methods. We note that the Matérn$\frac{3}{2}$ GP model can be efficiently implemented within the KF as a Markov process model [18]. Thus, it is possible to match the AUC of the KF approach and GP approach by replacing the near constant velocity model in the KF by the Markovianised Matérn model as described in [18]. However the results illustrate the significant improvement obtained using EVT as opposed to a simple gating mechanism based on the number of standard deviations between a datum and the expected position of the ship.

VI. Conclusion

Extreme value theory has proven to be an extremely successful framework for anomaly detection. Unlike novelty detection based directly on the sample distribution, extreme value distributions capture our beliefs that extreme events should become more extreme if large numbers of measurements are expected and vice versa. Such detection, however, has to be dynamic, context sensitive and timely if it is to be useful for marine tracking. Extreme value distributions alone are not readily adapted to perform this task.

In this paper we present an alternative to endowing extreme value distributions directly with dynamic properties. Our approach simultaneously models the dynamic properties of the underlying extreme value generative distribution and the dynamic properties of the data sampling process. To our knowledge this is the first time that extreme value distributions have been made dynamic through the use of Gaussian Processes.

Our approach offers several advantages. Gaussian Processes provide a flexible, non-parametric and intuitive tool to describe typical vessel dynamics. Also, measurement and prediction is performed in continuous time thus allowing on-demand anomaly detection. The sequential update of the Gaussian Process covariance matrix bypasses the need for inverting massive matrices and substantially reduces the computational burdens for which GPs are well known.

Our empirical experiments on vessel data suggest that the method is capable of detecting anomalies that resemble mooring or drifting, and unexpected departures from regular movements. The sample size prediction plays the important role of adapting the observation process in time. As the effective sample size reduces, the extreme value distribution approaches the regular Gaussian distribution, as Equation 15 suggests. With increasing density of observations, however, the extreme value distribution diverges and EVT bound increases. Although, the choice of Gaussian Process kernel function becomes less critical with increasing amounts of data, for smaller sample sizes, the kernel function is critical and our empirical results have shown that the Matérn$\frac{3}{2}$ kernel outperforms the near constant velocity model.

VII. Future work

The representative choice of distance as the dependent variable feature for anomaly detection is open to discussion. While it provides a single dimension and, thus, fast estimation it does fail to capture some aspects of ship tracks. To capture such features the GPS coordinates can be simultaneously modelled with a bivariate Gaussian Process and extrema modelling.

The training using typical vessel tracks will be extended to shipping lanes and vessel types. This allows anomaly detection not just on the basis of individual points but entire tracks and so offers the possibility of preventing accidents such as that of the MS Costa Concordia early in 2012. Kernel-regression based prediction of the sample size can be readily extended using Poisson Processes.

VIII. Acknowledgements

This work was funded by ISSG (In Service Support Group), Babcock Marine & Technology Division, Devonport Royal Dockyard.

References

[1] F. E. Grubbs, "Procedures for Detecting Outlying Observations in Samples," *Technometrics*, vol. 11, pp. 1–21, 1969.

[2] B. J. Rhodes, N. A. Bomberger, M. Seibert, and A. M. Waxman, "Maritime Situation Monitoring And Awareness Using Learning Mechanisms," in *Military Communications Conference*, vol. 1, 2005, pp. 646–652.

[3] S. Mascaro, A. E. Nicholson, and K. B. Korb, "Anomaly Detection in Vessel Tracks using Bayesian networks," in *Eighth UAI Bayesian Modeling Applications Workshop*, 2011, pp. 99–107.

[4] X. Li, J. Han, and S. Kim, "Motion-Alert: Automatic Anomaly Detection in Massive Moving Objects," in *IEEE Intelligence and Security Informatics*, 2006.

[5] J. Will, L. Peel, and C. Claxton, "Fast Maritime Anomaly Detection using Kd-Tree Gaussian Processes," in *IMA Maths in Defence Conference*, 2011.

[6] K. Laws, J. Vesecky, and J. Paduan, "Monitoring Coastal Vessels for Environmental Applications: Application of Kalman Filtering," in *10th Current, Waves and Turbulence Measurements (CWTM)*, 2011, pp. 39–46.

[7] R. Laxhammar, "Anomaly detection for sea surveillance," in *11th International Conference on Information Fusion*, 2008, pp. 1–8.

[8] R. O. Lane, D. A. Nevell, S. D. Hayward, and T. W. Beaney, "Maritime anomaly detection and threat assessment," in *13th Conference on Information Fusion (FUSION)*, 2010, pp. 1–8.

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

[10] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. the MIT Press, 2006.

[11] M. Osborne, "Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature," Ph.D. dissertation, University of Oxford, 2010.

[12] S. J. Roberts, "Extreme Value Statistics For Novelty Detection In Biomedical Signal Processing," in *Advances in Medical Signal and Information Processing*, 2000, pp. 166–172.

[13] H. Lee and S. J. Roberts, "On-Line Novelty Detection Using the Kalman Filter and Extreme Value Theory," in *19th International Conference on Pattern Recognition*, 2008, pp. 1–4.

[14] S. Coles, *An Introduction to Statistical Modelling of Extreme Values*. Springer Series in Statistics, 2001.

[15] S. L. Miller, W. M. Miller, and P. J. McWhorter, "Extremal dynamics: A unifying physical explanation of fractals, 1/f noise, and activated processes," *Journal of Applied Physics*, vol. 73, p. 6, 1992.

[16] M. Markou and S. Singh, "Novelty Detection: A Review Part 1: Statistical Approaches," *Signal Processing*, vol. 83, pp. 2481–2497, 2003.

[17] J. George, J. L. Crassidis, T. Singh, and A. M. Fosbury, "Anomaly Detection using Content-Aided Target Tracking," *Journal of Advances in Information Fusion*, vol. 6, pp. 39–56, 2011.

[18] J. Hartikainen and S. Särkkä, "Kalman Filtering and Smoothing Solutions to Temporal Gaussian Process Regression Models," in *Proc of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2010.