Adaptive Out-of-Control Point Pattern Detection in Sequential Random Finite Set Observations

Konstantinos Bourazas†, Savvas Papaioannou, and Panayiotis Kolios

***Abstract*— In this work we introduce a novel adaptive anomaly detection framework specifically designed for monitoring sequential random finite set (RFS) observations. Our approach effectively distinguishes between in-control data (normal) and out-of-control data (anomalies) by detecting deviations from the expected statistical behavior of the process. The primary contributions of this study include the development of an innovative RFS-based framework that not only learns the normal behavior of the data-generating process online but also dynamically adapts to behavioral shifts to accurately identify abnormal point patterns. To achieve this, we introduce a new class of RFS-based posterior distributions, named Power Discounting Posteriors (PD), which facilitate adaptation to systematic changes in data while enabling anomaly detection of point pattern data through a novel predictive posterior density function. The effectiveness of the proposed approach is demonstrated by extensive qualitative and quantitative simulation experiments.**

I. Introduction

Anomaly detection comprises techniques aimed at monitoring various processes, such as manufacturing, finance, and cybersecurity, to identify deviations from expected behavior that may indicate problems or inefficiencies. Its broad applicability spans fraud detection (e.g., credit cards, insurance, healthcare), cybersecurity, and fault diagnosis in critical systems, due to the valuable insights anomalies often provide. For instance, unusual network traffic may signal security breaches [1], abnormal medical findings may indicate malignancy [2], and atypical transactions may reveal financial fraud [3]. Despite decades of research and maturity in some domains, detecting anomalies remains challenging, particularly in stochastic point pattern data [4], where both the number and nature of observations are random. Most existing statistical [5], [6] and data-driven [7] methods assume fixed-size input structures (e.g., vectors, matrices), limiting their effectiveness in such settings.

Point patterns, however, consist of random sets of random vectors, with both spatial configuration and cardinality varying across instances. This variability renders traditional methods largely ineffective. To address this, Random Finite Set (RFS) theory [8], [9] offers a robust framework for handling random set-valued data and parameters. RFS is particularly suited for inference tasks involving an unknown and varying number of observations, with applications in state estimation [10], [11], [12], tracking [13], [14], [15], robotics [16], [17], [18], and surveillance [19], [20], [21].

However, its application in anomaly detection has been largely overlooked to date. More related to our work are the works in [22], [23], [24] and [25]. Specifically, the authors in [22] study the problem of model-based learning for classification, anomaly detection, and clustering of point pattern data. The approaches proposed in [22], however, are offline (i.e., they are not designed for observations obtained sequentially), and require a two-stage procedure: in the first stage, a training/calibration process is undertaken to learn the model's parameters, and only then can the respective task (i.e., anomaly detection) be performed. Similarly, offline two-stage RFS-based anomaly detection algorithms are proposed in [23] for hazard detection at construction sites, and in [24] for detecting anomalies in industrial manufacturing. The problem of anomaly detection with RFS observations is also investigated in [25], where the authors propose a methodology for detecting defects in industrial settings using RFS energy-based models. However, this methodology is not adaptive; the RFS energy-based model is known a priori, and its parameters, learned using offline calibration data, remain fixed thereafter. In contrast, our proposed approach operates within the Bayesian framework, where the model parameters are themselves treated as random variables following (non-informative) conjugate prior distributions, and therefore can be adapted to match the normal behavior of the generating process.

Specifically, we propose a novel adaptive anomaly detection framework for sequential random finite set (RFS) observations. We assume an unknown Poisson RFS process generates sequential point pattern data. Our system learns and monitors the process in an online fashion while determining whether the data remains in statistical control [26]. Data conforming to the expected behavior are considered In-Control (IC), or normal, while deviations signify Out-Of-Control (OOC) data, or anomalies. Anomalies indicate either a temporary deviation from the in-control process behavior or a fundamental shift, necessitating an update to the established process statistical state. Our contributions are the following:

- We investigate the problem of anomaly detection with sequential random finite set (RFS) observations, and propose an adaptive anomaly detection framework. This framework learns the process's normal behavior in an on-line fashion, adapting to behavioral shifts, while being capable of detecting out-of-control (OOC) point pattern data (i.e., anomalies) that do not conform to the

The authors are with the †Department of Economics, Athens University of Economics and Business, Athens, Greece, and the KIOS Research and Innovation Centre of Excellence (KIOS CoE), University of Cyprus, Nicosia, 1678, Cyprus. E-mail : kbourazas@aueb.gr, {papaioannou.savvas, pkolios }@ucy.ac.cy. This work is implemented under the Border Management and Visa Policy Instrument (BMVI) and is co-financed by the European Union and the Republic of Cyprus (GA:BMVI/2021-2022/SA/1.2.1/015), and also supported by the European Union's H2020 research and innovation programme under grant agreement No 739551 (KIOS CoE - TEAMING) and through the Deputy Ministry of Research, Innovation and Digital Policy of the Republic of Cyprus.

expected behavior of the process.
- We introduce the Power Discounting Posteriors (PD), a new class of RFS-based posterior distributions designed for on-line learning and process monitoring. Subsequently, we derive a Bayesian RFS predictive check statistic, allowing robust detection of OOC data points, or anomalies.
- Finally, the effectiveness of the proposed approach is demonstrated through extensive qualitative and quantitative simulation experiments.

The paper is organized as follows. In Section II, we provide the background on the theory of random finite sets. In Section III, we formulate the problem tackled in this work, and subsequently, in Section IV, we discuss the details of the proposed approach. Finally, in Section V, we evaluate the proposed approach, and in Section VI, we conclude the paper.

II. Preliminaries

A. Random Finite Sets

A random finite set (RFS) is a finite-set-valued random variable where both the number of elements in the set and the values of these elements are random. Therefore, the main difference between a random finite set and a random vector lies in two aspects for the former: firstly, the number of elements, denoted as n, is itself random, and secondly, these elements are not only random but also unordered.

More specifically, an RFS X is completely specified by: a) its cardinality distribution $\rho(n) = p(|X| = n),\ n \in \mathbb{N}$, which defines the probability distribution over the number of elements in X, and b) by a family of conditional joint probability distributions $p(x_1, \ldots, x_n|n)$ that characterize the distribution (i.e., spatial density) of its elements $x_1, \ldots, x_n \in \mathcal{X}$ over the state space $\mathcal{X}$. The probability density function (pdf) $f(X) = f(\{x_1, \ldots, x_n\})$ of the RFS X, with $f : \mathcal{F}(\mathcal{X}) \to [0, \infty)$ where $\mathcal{F}(\mathcal{X})$ is the space of all finite subsets of $\mathcal{X}$, is given by:

$$f(\{x_1, \ldots, x_n\}) = \rho(n) \sum_{\varpi} p(x_{\varpi(1)}, \ldots, x_{\varpi(n)}|n) \quad (1)$$

where ϖ is the permutation of the elements $\{1, \ldots, n\}$, and is used as shown above to account for the fact that the elements in X are unordered. Moreover, for joint symmetric conditional densities, Eq. (1) simplifies to:

$$f(\{x_1, \ldots, x_n\}) = \rho(n)n!p(x_1, \ldots, x_n|n) \quad (2)$$

where $n!$ denotes the n-factorial, and $p(x_1, \ldots, x_n|n)$ is a symmetric density, meaning that its value remains unchanged for all of the possible $n!$ permutations of its input variables. The notion of integration is then given by the set-integral, which is defined as:

$$\int f(X)dX = f(\emptyset) + \sum_{n=1}^{\infty} \frac{1}{n!} \int f(\{x_1, \ldots, x_n\})\, dx_1 \ldots dx_n \quad (3)$$

where by convention $f(\emptyset) = \rho(0)$. Finally, it is straightforward to show that when the elements $x \in X$ of the RFS X are independent and identically distributed (iid) according to the probability density $p(x)$ on $\mathcal{X}$, the pdf of X is given by:

$$f(X) = \rho(n)n! \prod_{x \in X} p(x) \quad (4)$$

Additionally, when the cardinality n follows a Poisson distribution with parameter λ, refereed to as Pois(λ), i.e., the cardinality distribution is given by $\rho(n) = \frac{e^{-\lambda}\lambda^n}{n!}$, the RFS X becomes a Poisson RFS with density given by:

$$f(X) = e^{-\lambda} \prod_{x \in X} \kappa(x) \quad (5)$$

where $\kappa(x) = \lambda p(x)$ is called the intensity function of X, which when integrated over any closed subset $S \subseteq \mathcal{X}$ gives the expected number of elements $\mathbb{E}[n]$ in S i.e., $\mathbb{E}[n] = \int_S \kappa(x)dx$. The Poisson RFS, due to its importance in diverse application scenarios [27], serves as the primary focus of this work for anomaly detection. However, the proposed approach can be generalized in other RFS models. It is also worth mentioning that the non-uniformity of the reference measure in the RFS framework causes problems in anomaly detection [22]. To overcome this problem, the ranking function has been proposed [22], which for the iid cluster model is given by:

$$r(X) \propto \rho(n) \frac{\prod_{x \in X} p(x)}{(\|p(x)\|_2^2)^n}, \quad (6)$$

where $\|\cdot\|_2$ is the L_2 norm.

III. Problem Formulation

We consider a discrete-time, real valued, and bounded Poisson RFS process with density function governed by Eq. (5). This process generates sequentially RFS observations i.e., at each time-step t it generates the RFS observation $X_t = \{x_{1,t}, \ldots, x_{n,t}\}$, where $n_t = |X_t| \sim$ Pois(λ_t), and $x_{i,t} \sim N_d(\mu_t, \Sigma_t), i \in \{1, \ldots, n_t\}$ i.e., the features or elements are drawn from a d-variate Normal distribution with mean vector μ_t and covariance matrix Σ_t. The problem tackled in this work can be stated as follows as:

Given sequential Poisson RFS observations $\{X_t|t > 0\}$, our objective is to dynamically learn the posterior predictive density $f(X_{t+1}|X_{1:t})$ for the subsequent time-step $t + 1$, based on all preceding observations $X_{1:t}$ up to time t. Following this, we aim to assess the "extremeness" in terms of the statistical significance for the newly received RFS observation X_{t+1} in relation to the process' expected behavior, and detect OOC point pattern data (i.e., anomalies) that do not conform to the process' anticipated behavior.

A high-level overview of the proposed framework is given next and detailed discussion in Sec. IV. Within the Bayesian framework, we consider the rate parameter λ_t along with the mean vector μ_t and the covariance matrix Σ_t as unknown. We denote the list of the unknown parameters as Θ_t i.e., $\Theta_t = (\lambda_t, \mu_t, \Sigma_t)$, that belong in a parametric space $\boldsymbol{\Theta} = \mathbb{R}^+ \times \mathbb{R}^d \times \mathbb{R}^{d \times d}$. By making use Eqs. (4)-(5), in this work the likelihood function $f(X_t|\Theta_t)$ of the Poisson RFS process

can be computed as:

$$f(X_t|\Theta_t) = e^{-\lambda_t} \prod_{j=1}^{n_t} \lambda_t p(x_{j,t}|\mu_t, \Sigma_t) \tag{7}$$

Our goal is to learn on-line the parameters Θ_t in Eq. (7) through sequential RFS observations, aiming to ascertain the anticipated behavior of the process. Simultaneously, we seek to identify observations that deviate from this expected behavior (i.e., anomalies), by taking into account the uncertainty of Θ_t. We formulate this uncertainty assuming that Θ_t follows a prior distribution, denoted as $\pi(\Theta_t)$. The prior distribution represents our belief on the parameters, with the non-informative priors [28] being a plausible choice in cases of process ignorance. Given the RFS observations $X_{1:t}$ up to time-step t, we compute the posterior density, denoted as $\pi(\Theta_t|X_{1:t})$, which is given by:

$$\pi\left(\Theta_t|X_{1:t}\right) \propto f\left(X_{1:t}|\Theta_t\right)\pi(\Theta_t), \tag{8}$$

resulting in an updated version of the prior informed by the likelihood. Extending the framework of the classical posterior in the Bayesian approach, we introduce the Power Discounting (PD) posteriors $\pi(\Theta_t|X_{1:t}, \alpha_0)$, further discussed in Sec. IV-A. These allow for adaptation to changes in the process's behavior by learning the parameters Θ_t over time using a discounting factor α_0. Subsequently, the posterior predictive distribution $f(X_{t+1}|X_{1:t}, \alpha_0)$, is given by:

$$f(X_{t+1}|X_{1:t}, \alpha_0) = \int f(X_{t+1}|\Theta_t)\pi(\Theta_t|X_{1:t}, \alpha_0)d\Theta_t, \tag{9}$$

which derived by integrating out the unknown parameters with respect to the posterior. Based on this, to determine whether the received RFS observation X_{t+1} at time-step $t+1$ conforms to the process's expected behavior (i.e., is normal or anomalous), we leverage the independence between the cardinality distribution $\rho(n_{t+1})$ and the conditional spatial density $p(x_{1,t+1}, \ldots, x_{n_{t+1},t+1}|n_{t+1})$, to express the posterior in closed form as $\pi(\Theta_t|X_{1:t}, \alpha_0) = \pi(\lambda_t|n_{1:t}, \alpha_0)\pi(\mu_t, \Sigma_t|X_{1:t}, \alpha_0)$. Thus, we can derive the posterior predictive distributions of the cardinality, and the spatial density as shown in Eq. (10), and Eq. (11) respectively:

$$\rho\left(n_{t+1}|n_{1:t}, \alpha_0\right) = \int \rho(n_{t+1}|\lambda_t)\pi(\lambda_t|n_{1:t}, \alpha_0)d\lambda_t, \tag{10}$$

$$\begin{aligned} p\left(\bar{x}_{t+1}|X_{1:t}, \alpha_0\right) = \int\int p(\bar{x}_{t+1}|\mu_t, \Sigma_t)\times \\ \times\, \pi(\mu_t, \Sigma_t|X_{1:t}, \alpha_0)d\mu_t d\Sigma_t, \end{aligned} \tag{11}$$

where $\bar{x}_{t+1} = \sum_{j=1}^{n_{t+1}} x_{j,t+1}/n_{t+1}$ is the sufficient statistic of the mean vector summarizing all the available information from the features. This derivation is presented in Sec. IV-B.

Subsequently, we employ posterior predictive checks to assess the statistical significance of a new observation relative to the observed sequence of observations. Specifically, we define two posterior probabilities, denoted as pr^n_{t+1} for the cardinality, and $pr^{x|n}_{t+1}$ for the features given cardinality at time $t+1$, which will play the role of posterior predictive p-values [29], measuring how extreme the observation is for the posterior prediction.

Finally, we assess the presence of an anomaly via the Fisher's combined probability test [30] which in this work can be defined as:

$$P_{t+1} = -2\log\left(pr^n_{t+1} \cdot pr^{x|n}_{t+1}\right). \tag{12}$$

For uniformly distributed probabilities pr^n_{t+1} and $pr^{x|n}_{t+1}$ (discussed in detail in Sec. IV-B), the distribution of P_{t+1} under the null hypothesis of non-anomaly follows a chi-squared distribution with 4 degrees of freedom, i.e., $P_{t+1} \sim \mathcal{X}^2_4$. Thus, we raise an alarm if:

$$P_{t+1} > q(1-\alpha), \tag{13}$$

where $q(\cdot)$ is the quantile function of $\mathcal{X}^2_4$ distribution, and α is a predetermined false alarm rate α.

IV. Adaptive Anomaly Detection in Sequential RFS Observations

A. Power Discounting posteriors for RFS Point-pattern Data

In a sequential process, as the observation horizon grows, the posterior distribution becomes more informative. While beneficial in some cases, this can hinder adaptation to small systematic changes. Accumulated evidence creates an "inertia" effect, where posterior parameters require many observations to shift, making the distribution increasingly resistant to change. For this reason, we introduce a new class of posterior distributions, the Power Discounting posteriors (PD). PD posteriors, via a discounting factor α_0, weigh the importance of the received observations, giving more weight to the most recent. The idea of power discounting is not new in Bayesian statistics but in a different set-up [31], [32]. PD posteriors differentiate, as they implement sequential power discounting; they are suitable for monitoring, but at the same time, they leverage the temporal significance of the observations to facilitate adaptation to systematic changes, and improve the anomaly detection capability. For the unknown parameters Θ_t, the general form of the PD posterior at time time t is:

$$\pi\left(\Theta_t|X_{1:t}, \alpha_0\right) \propto f\left(X_t|\Theta_t\right)\prod_{i=1}^{t-1} f\left(X_{1:t-1}|\Theta_t\right)^{\alpha_0^{t-i}}\pi(\Theta_t)^{\alpha_0^t}, \tag{14}$$

where the parameter $0 \leq \alpha_0 \leq 1$ is the discounting factor weighting the influence of the observations in the posterior. In extreme cases, such as when $\alpha_0 = 0$, the process transforms into pure adaptation, as only the latest observation impacts the posterior while ignoring the previous observation. Conversely, when $\alpha_0 = 1$, then equal weight is given to all observations, making the posterior more informative and appropriate for history-based monitoring.

Because of the independence between λ_t and (μ_t, Σ_t), their joint prior can be written as $\pi\left(\lambda_t, \mu_t, \Sigma_t\right) = \pi\left(\lambda_t\right)\pi\left(\mu_t, \Sigma_t\right)$ (this product keeps for the posterior as well, i.e., $\pi(\lambda_t, \mu_t, \Sigma_t|X_{1:t}, \alpha_0) = \pi(\lambda_t|n_{1:t}, \alpha_0)\pi(\mu_t, \Sigma_t|X_{1:t}, \alpha_0)$).

Lemma 1 and Lemma 2 derive the posterior distribution of λ_t and the joint posterior distribution of (μ_t, Σ_t), respectively.

Lemma 1: Assuming the prior $\lambda_t \sim G(c_0, d_0)$, i.e., a Gamma distribution with hyperparameters c_0, and d_0, then the PD posterior distribution is conjugate $\lambda_t|(n_{1:t}, \alpha_0) \sim G(c_t, d_t)$, with c_t, and d_t, the updated parameters.

Proof: At time t, the posterior of the rate is:

$$\pi(\lambda_t|n_{1:t}, \alpha_0) \propto \rho(n_t|\lambda_t) \prod_{i=1}^{t-1} \rho(n_i|\lambda_t)^{\alpha_0^{t-i}} \pi(\lambda_t)^{\alpha_0^t} \quad (15)$$

Keeping only any expression that contains the unknown parameters we have that $\pi(\lambda_t|n_{1:t}, \alpha_0) \propto$:

$$e^{-\lambda_t} \lambda_t^{n_t} \prod_{i=1}^{t-1} \left(e^{-\lambda_t} \lambda_t^{n_i}\right)^{\alpha_0^{t-i}} \left(e^{-d_0\lambda_t} \lambda_t^{c_0}\right)^{\alpha_0^t} \implies$$
$$\pi(\lambda_t|n_{1:t}, \alpha_0) \propto e^{-\lambda_t d_t} \lambda_t^{c_t},$$

where $c_t = \alpha_0^t c_0 + n_t + \sum_{i=1}^{t-1} \alpha_0^{t-i} n_i$ and $d_t = \alpha_0^t d_0 + 1 + \sum_{i=1}^{t-1} \alpha_0^{t-i}$. Thus, $\lambda_t|(n_t, \alpha_0) \sim G(c_t, d_t)$. ∎

Lemma 2: Assuming the prior $(\mu_t, \Sigma_t) \sim NIW(m_0, l_0, \nu_0, \Psi_0)$, i.e., a Normal-Inverse Wishart with hyperparameters m_0, l_0, ν_0, and Ψ_0, then the PD posterior distribution is conjugate $(\mu_t, \Sigma_t)\,|\,(X_{1:t}, \alpha_0) \sim NIW(m_t, l_t, \nu_t, \Psi_t)$, with m_t, l_t, ν_t, and Ψ_t, the updated parameters.

Proof: At time t, the joint posterior of the mean vector and the covariance matrix is:

$$\pi(\mu_t, \Sigma_t|X_{1:t}, \alpha_0) \propto \prod_{j=1}^{n_t} p(x_{j,t}|\mu_t, \Sigma_t) \prod_{i=1}^{t-1} \left(\prod_{j=1}^{n_i} p(x_{j,i}|\mu_t, \Sigma_t)\right)^{\alpha_0^{t-i}} \times \pi(\mu_t, \Sigma_t)^{\alpha_0^t} \quad (16)$$

Keeping only any expression that contains the unknown parameters we have that $\pi(\mu_t, \Sigma_t|X_{1:t}, \alpha_0) \propto$

$$|\Sigma_t|^{-\sum_{i=1}^{t} \alpha_0^{t-i} n_i/2} \exp\left\{-\frac{1}{2} \operatorname{tr} \Sigma_t^{-1} \sum_{j=1}^{n_i} x_{j,i} x_{j,i}^T\right\} \times \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{t} \alpha_0^{t-i} n_i - 2 \sum_{i=1}^{t} \alpha_0^{t-i} \sum_{j=1}^{n_i} x_{j,i}^T \Sigma_t^T \mu_t\right)\right\}$$
$$\times |\Sigma_t|^{-\sum_{i=1}^{t-1} \alpha_0^{t-i} n_i/2} \exp\left\{-\frac{1}{2} \operatorname{tr} \Sigma_t^{-1} \sum_{j=1}^{n_i} x_{j,i} x_{j,i}^T\right\} \times \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{t-1} \alpha_0^{t-i} n_i - 2 \sum_{i=1}^{t-1} \alpha_0^{t-i} \sum_{j=1}^{n_i} x_{j,i}^T \Sigma_t^T \mu_t\right)\right\}$$
$$\times |\Sigma_t|^{-\alpha_0^t(\nu_0+d+2)/2} \exp\left\{-\frac{1}{2} \alpha_0^t \operatorname{tr} \Sigma_t^{-1} \Psi_0\right\} \times \exp\left\{-\frac{\alpha_0^t l_0}{2} (\mu_t - m_0)^T \Sigma_t^{-1} (\mu_t - m_0)\right\}$$

By doing the appropriate operations and completing the quadratic forms we have:

$$\pi(\mu_t, \Sigma_t|X_{1:t}, \alpha_0) \propto |\Sigma_t|^{-(\nu_t+d+2)/2} \exp\left\{-\frac{1}{2} \operatorname{tr} \Sigma_t^{-1} \Psi_t\right\} \times \exp\left\{-\frac{\lambda_t}{2} (\mu_t - m_t)^T \Sigma_t^{-1} (\mu_t - m_t)\right\}$$

where $m_t = \dfrac{\alpha_0^t l_0 m_0 + \sum_{j=1}^{n_t} x_{j,t} + \sum_{i=1}^{t-1} \alpha_0^{t-i} \sum_{i=1}^{n_i} x_{j,i}}{\alpha_0^t l_0 + n_t + \sum_{i=1}^{t-1} n_i \alpha_0^{t-i}}$,

$l_t = \alpha_0^t l_0 + n_t + \sum_{i=1}^{t-1} n_i \alpha_0^{t-i}$, $\nu_t = \alpha_0^t \nu_0 + n_t + \sum_{i=1}^{t-1} n_i \alpha_0^{t-i}$

and

$$\Psi_t = \alpha_0^t \Psi_0 + \alpha_0^t l_0 m_0 m_0^T + \sum_{j=1}^{n_t} x_{j,t} x_{j,t}^T + \sum_{i=1}^{t-1} \alpha_0^{t-i} \sum_{j=1}^{n_i} x_{j,i} x_{j,i}^T - \frac{1}{\alpha_0^t l_0 + n_t + \sum_{i=1}^{t-1} n_i \alpha_0^{t-i}} \times \left(\alpha_0^t l_0 m_0 + \sum_{j=1}^{n_t} x_{j,t} + \sum_{i=1}^{t-1} \alpha_0^{t-i} \sum_{j=1}^{n_i} x_{j,i}\right) \times \left(\alpha_0^t l_0 m_0 + \sum_{j=1}^{n_t} x_{j,t} + \sum_{i=1}^{t-1} \alpha_0^{t-i} \sum_{j=1}^{n_i} x_{j,i}\right)^T.$$

Thus, $(\mu_t, \Sigma_t)\,|\,(X_{1:t}, \alpha_0) \sim NIW(m_t, l_t, \nu_t, \Psi_t)$. ∎

In cases where prior knowledge is lacking, it is important to consider non-informative priors. Among these, we recommend the use of the Jeffreys' prior [33]. Its density function is proportional to the square root of the determinant of the Fisher information matrix, i.e., $\pi(\Theta) \propto |\mathcal{I}(\Theta)|^{1/2}$, while, a property, which is of key importance, is its invariance under a change of coordinates for the unknown parameters. In our set-up, we have $\pi(\lambda_t) \propto \lambda_t^{-1/2}$ and $\pi(\mu_t, \Sigma_t) \propto |\Sigma_t|^{-(d+2)/2}$ for the rate and the features' parameters, respectively. The total weight of information the posterior conveys to the observed sets arises from a geometric series, and specifically, when $t \to +\infty$, then for cardinality the total weight is of $1/(1-\alpha_0)$, and for the features the expected weight is for cardinality the total weight is of $\lambda_t/(1-\alpha_0)$. An illustrative example of the proposed PD posterior for monitoring the cardinality distribution $\rho(n)$ is depicted in Fig. 1. Specifically, the figures shows the cardinality of the received RFS observations over time, with the corresponding posterior mean shown in blue. The initial 50 observations are simulated from a Pois(10) distribution, followed by a smooth increase in the rate parameter λ_t to 12 for the subsequent 30 observations. Finally, the rate λ_t smoothly decreases to 5 for the last 20 observations. To assess the effect of the discounting factor we set three values for α_0, and specifically $\alpha_0 \in \{0.8, 0.9, 1\}$, while a non-informative prior is employed. We observe that the smaller the α_0, the more the posterior mean adapts to changes.

B. Detecting Anomalies in Sequential RFS Observations

The proposed methodology for anomaly detection is based on the posterior predictive distribution, which will be the

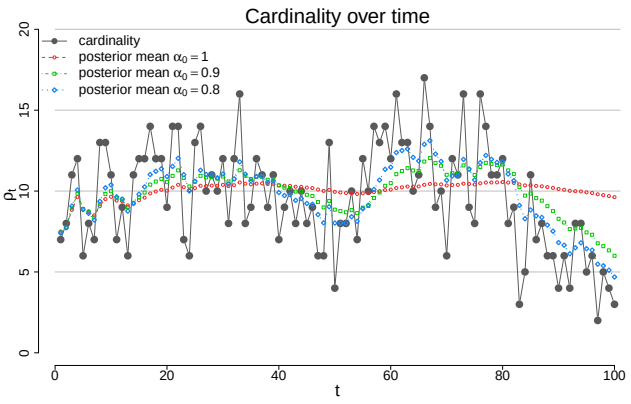

Fig. 1. The figure illustrates the adaptive nature of the proposed approach i.e., how the propped approach tracks the cardinality, and its posterior mean in time for different values of $\alpha_0 \in \{0.8, 0.9, 1\}$.

representative of the likelihood, taking into account the uncertainty for $\Theta_t = (\lambda_t, \mu_t, \Sigma_t)$. In our set-up, the general form of predictive density of X_{t+1} is derived by integrating out the unknown parameters with respect to the posterior $p(\Theta_t|X_{1:t}, \alpha_0)$. Specifically:

$$f(X_{t+1}|X_{1:t}, \alpha_0) = \int_{\Theta_t} f(X_{t+1}|\Theta_t)\pi(\Theta_t|X_{1:t}, \alpha_0)d\Theta_t \quad (17)$$

It is well known the non-uniformity issues of $f(X_{t+1}|\Theta_t)$ (and transition to $f(X_{t+1}|X_{1:t}, \alpha_0)$), which fails to peak at the most probable values. The Ranking Function (RF) was introduced in [22] to alleviate these issues arising from the likelihood, but we tackle this problem in a different way. Specifically, considering the independence between the cardinality and the features given the cardinality, we will employ two independent predictive checks. These checks will assess anomalies for the cardinality and the features, respectively, based on their corresponding predictive distributions. Then, we will combine the anomaly evidence of the two predictive checks into the Fisher score in Eq. (12) for assessing the presence of anomaly in X_{t+1} via (13). Starting from n_{t+1}, its posterior predictive distribution [34] is a Negative Binomial:

$$n_{t+1}|(n_{1:t}, \alpha_0) \sim NB\left(c_t, \frac{d_t}{d_t + 1}\right). \quad (18)$$

To employ the predictive check for the cardinality, we define the set of Highest Predictive Probabilities (HPrP) as

$$\mathcal{R}(n_{t+1}) = \{n : \rho(n|n_{1:t}, \alpha_0) > \rho(n_{t+1}|n_{1:t}, \alpha_0)\} \quad (19)$$

which is the set with the values of the posterior predictive with higher probability than the observed cardinality n_{t+1}, with $\mathcal{R}(n_{t+1}) = \emptyset$ if n_{t+1} is the mode of the posterior predictive density. Note that HPrP is closely related to the Highest Predictive Mass (HPrM), introduced in [35] for a Bayesian approach in online anomaly detection for univariate processes. However, in our set-up, we need to define the probability that quantifies the discrepancy between the observed cardinality and its posterior predictive distribution and combine this probability with the corresponding probability of the features rather than employ a decision rule for the cardinality. Thus, we define the probability of obtaining results at least as extreme as n_{t+1} by

$$pr^n_{t+1} = \sum_{n \notin \mathcal{R}(n_{t+1})} \rho(n|n_{1:t}, \alpha_0). \quad (20)$$

The choice of HPrP in (20) is optimal in the sense that minimizes the discrete measure $m(\mathcal{R}^c) = \sum_i \delta_{n_i} (\rho(n_i|n_{1:t}, \alpha_0) > \rho(n_{t+1}|n_{1:t}, \alpha_0))$, where δ_{n_i} is the Delta Dirac function. In other words, HPrP is the shortest region that achieves the sum $\sum_{n \in \mathcal{R}(n_{t+1})} \rho(n|n_{1:t}, \alpha_0)$, and consequently maximizes the region that indicates a discrepancy, an extremely useful property for anomaly detection. Essentially, pr^n_{t+1} aggregates all the probabilities of values not belonging to $\mathcal{R}(n_{t+1})$, thus having smaller or equal probabilities than n_{t+1}. Large values for pr^n_{t+1} imply that set $\mathcal{R}(n_{t+1})$ has few values, and thus n_{t+1} has a relatively high probability in the distribution; small values are associated with extreme values in the posterior predictive distribution, essentially indicating an anomaly.

Regarding the distribution of pr^n_{t+1}, it is discrete in the range $[0, 1]$, as $\rho(\cdot|n_{1:t}, \alpha_0)$ is discrete. However, under certain conditions, the distribution of pr^n_{t+1} approximates asymptotically the standard uniform distribution $U(0, 1)$. Specifically, it is well known that as $c_t \to +\infty$ and the probability of "success" $d_t/(d_t + 1) \to 1$, i.e., when the posterior becomes informative, then the Negative Binomial approximates the Poisson with $\lambda_t = c_t/d_t$. For large values of λ_t, the $\mathcal{R}(n_{t+1})$ approximates the rejection region of a two-tailed test of Normal distribution, and pr^n_{t+1} approximates the classical two sided p-value, which follows a $U(0, 1)$. But even in cases where the conditions of the approximation are not met, the distribution is uniform as possible due to its discreteness. Furthermore, the approximation would be mainly poor for values close to one, rather than close to zero, which are of main interest in detecting an anomaly, due to the large number of values with small probabilities in $\rho(\cdot|n_{1:t}, \alpha_0)$. Continuing with the distribution of the features given the cardinality, we express an anomaly it terms of a disorder in the mean vector. From basic properties of the Normal distribution, it is known that the distribution of $\bar{x}_{t+1}$ is:

$$\bar{x}_{t+1} \sim N\left(\mu_t, \Sigma_t/n_{t+1}\right). \quad (21)$$

Using the posterior $\pi\left(\mu_t, \Sigma_t|X_{1:t}, \alpha_0\right)$, we derive the posterior predictive as in (17) for the mean vector which is [34]

$$\bar{x}_{t+1}|\left(X_{1:t}, \alpha_0\right) \sim t_{\nu_t - d + 1}\left(m_t, \frac{\lambda_t + 1}{\lambda_t(\nu_t - d + 1)n_{t+1}}\Psi_t\right),$$

i.e., a d-variate t-Student distribution, with degrees of freedom $\nu_t - d + 1$, mean vector m_t and covariance matrix $\frac{\lambda_t+1}{\lambda_t(\nu_t-d+1)n_{t+1}}\Psi_t$. The predictive check for the features is based on the Hotelling statistic [36], which is uses the Mahalanobis distance, and is admissible for testing anomalies in Normal data [37]. In our case, the Hotelling type statistic used will be the Mahalanobis distance between the observed sample mean vector and the posterior predictive mean vector, weighted by the posterior predictive covariance matrix, i.e.,

$$T^2_{t+1} = (\bar{x}_{t+1} - m_t)^T \left(\frac{(\lambda_t + 1)d}{\lambda_t(\nu_t - d + 1)n_{t+1}}\Psi_t\right)^{-1} (\bar{x}_{t+1} - m_t),$$

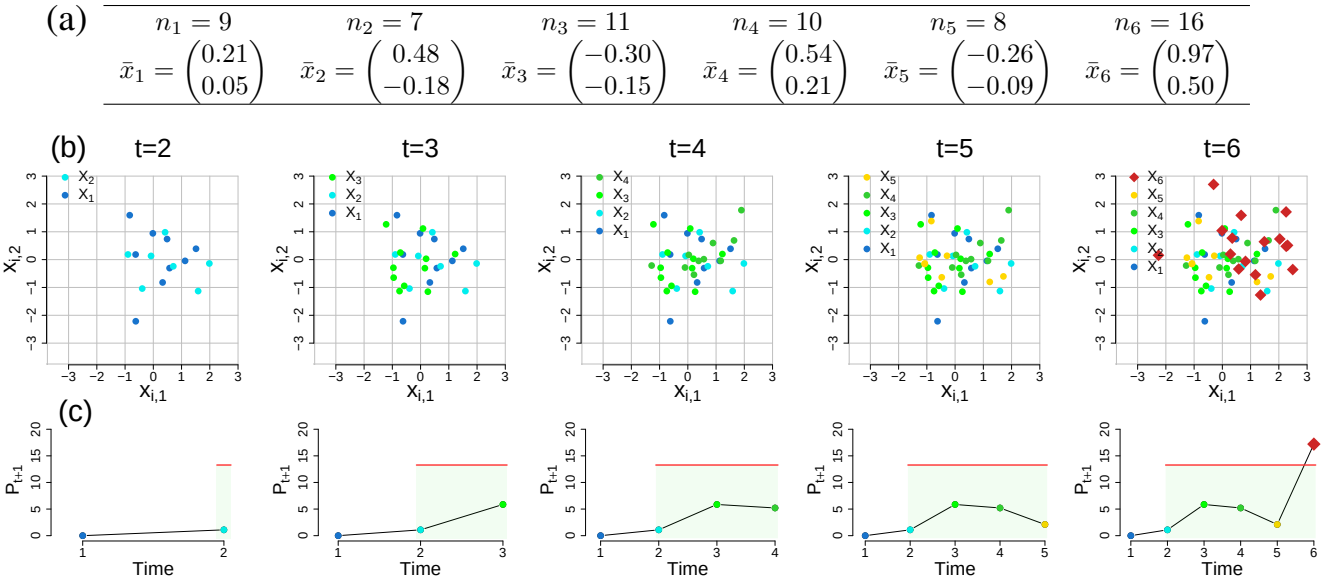

Fig. 2. The figure depicts an illustrative example of the proposed approach. The top panel (a) displays the cardinality and sample mean vector of the observed RFSs over time, the center panel (b) shows scatterplots of the RFSs, and at the lower panel (c) we provide the Fisher score P_{t+1} in Eq. (12) based on the predictive checks. The Jeffreys' prior is used, while $\alpha_0 = 1$ for the PD in Eq. (14), and $\alpha = 1/100$ for the quantile function function in Eq. (13). In (c), the red line is the decision limit $q(1-\alpha)$, while the light green area indicates the no-anomaly region for the process. An anomalous RFS observation (marked with red $\diamond$) is detected at time-step 6 as shown in the figure.

The distribution of T^2_{t+1} test statistic is an F with d and $\nu_t - d + 1$ degrees of freedom, i.e., $T^2_{t+1} \sim F_{d,\nu_t-d+1}$, with $\nu_t > d - 1$ [38]. Significantly large values of T^2_{t+1} indicate a large Mahalanobis distance, and consequently an potential anomaly. Thus, we define the probability of obtaining results at least as extreme as $\bar{x}_{t+1}$ by

$$pr^{x|n}_{t+1} = 1 - CF_{d,\nu_n-d+1}\left(T^2_{t+1}\right), \tag{22}$$

where $CF_{d,\nu_n-d+1}(\cdot)$ is the cumulative distribution function of the F distribution with degrees of freedom d and $\nu_n - d + 1$. It is straightforward to prove that $pr^{x|n}_{t+1} \sim U(0,1)$ under the assumption of no anomaly. Finally, to combine the information of the the predictive checks for the cardinality and the features given the cardinality, we substitute the uniformly distributed probabilities introduced in (20) and (22) into (12). Thus, we obtain the Fisher test statistic P_{t+1}, which summarizes the "extremeness" of the X_{t+1} under its posterior predictive distribution. We trigger an alarm based on (13), i.e., if $P_{t+1} > q(1-\alpha)$, where $q(\cdot)$ is the quantile function of $\mathcal{X}^2_4$ distribution, and α is a predetermined false alarm rate α. Figure 2 visualizes a paradigm with a simultaneous change in the cardinality and the mean vector of the features. Precisely, for the first five RFS we generate $n_i \sim \text{Pois}(10)$, $i \in \{1, \ldots, 5\}$, and $x_{j,i} \sim N_2(0_2, I_{2\times 2})$, i.e., a bivariate standard Normal distribution, where $j \in \{1, \ldots, n_i\}$. At the sixth RFS, we introduce an anomaly by increasing the rate parameter of the cardinality to 16 and adding one standard deviation to the first component of the mean vector and half a standard deviation to the second component, i.e., $x_{j,6} \sim N_2((1\ 0.5)^T, I_{2\times 2})$. The Jeffreys' prior is used, while we set the discounting factor $\alpha_0 = 1$ (i.e., no discounting) and the false alarm rate $\alpha = 1/100$. As we observed, the anomaly is detected at time $t = 6$, as X_6 deviates significantly from the previous for both the cardinality ($n_6 = 16$) and the mean vector ($\bar{x}_6 = (0.97\ 0.50)^T$).

V. Evaluation

In this section, we compare the performance of the proposed Bayesian predictive checks (PC) methodology against the ranking function (RF) introduced in [22] under various scenarios of anomalies in the sequence of Poisson RFS observations. Regarding the IC process, we assume batches of 30 RFSs where the cardinality follows a Pois(10), and the features follow the bivariate standard Normal distribution,

Regarding the IC process, we generate 10,000 batches of $T = 30$ RFSs where the cardinality follows a Pois(10), and the features follow the $N_2(0_2, I_{2\times 2})$, i.e. the bivariate standard Normal distribution. For the case of OOC data, we introduce anomalies in the IC sequences in each time-step $t \in \{2, \ldots, T\}$, drawing samples from the following 5 scenarios:

1) Spatial Density: we draw from a Normal distribution with mean vector $\mu'_t = (1\ 1)^T$, i.e., we introduce a shift size of 1 standard deviation for each component in the mean vector (Fig. 3(a)).
2) Cardinality Distribution: we draw samples from a Poisson with rate $\lambda' = 20$, i.e., we introduce an increase to the cardinality rate (Fig. 3(b)).
3) Cardinality Distribution: we draw samples from a Poisson with rate $\lambda' = 2$, i.e., we introduce a decrease to the cardinality rate (Fig. 3(c)).
4) Spatial Density and Cardinality: we draw samples from a Poisson with rate $\lambda' = 15$ and a Normal with mean vector $\mu'_t = (1\ 1)^T$, i.e., we introduce a moderate increase to the cardinality rate and a moderate shift for the mean vector (Fig. 3(d)).

5) Spatial Density and Cardinality: we draw from a Poisson with rate $\lambda' = 5$ and a Normal with mean vector $\mu'_t = (1\ 1)^T$, i.e., we introduce a moderate decrease to the cardinality rate and a moderate shift for the mean vector (Fig. 3(e)).

It is important to note that the 5 investigated OOC scenarios pose a significant challenge due to subtle deviations in cardinality distribution and spatial density of the features, relative to the IC process. Next, we demonstrate how our proposed approach handles these challenging scenarios compared to existing state-of-the-art methods.

PC requires the definition of a prior distribution, so, within this simulation study, we will take the opportunity to examine its sensitivity performance under the absence or the presence of prior information. Precisely, we will use the non-informative Jeffreys' prior (denoted as J) while for the informative prior setting (denoted as inf) we assume $\lambda \sim G(50.5, 5)$ and $(\mu_t, \Sigma_t) \sim N(m_0 = 0_2, l_0 = 50, \nu_0 = 48, \Psi_0 = 49 \cdot I_{2\times 2})$, where 0_2 is the zero vector and $I_{2\times 2}$ is the diagonal matrix. Note that this is the resulting posterior on average for a process with five IC RFSs using the Jeffreys' prior and setting $\alpha_0 = 1$. Furthermore, to assess the effect of the discounting parameter α_0 to the performance, we set $\alpha_0 \in \{0.8, 0.9, 1\}$. Thus, we have $2 \times 3 = 6$ versions of PC for the different choices of the prior and α_0. Regarding the competing method RF (shown as a dotted black line in Fig. 3), we use the theoretical values of the IC process to achieve the best possible performance. In other words, for the RF method we assume that the parameters are completely known (not estimated from a training dataset), while for PC are estimated by the information of the prior (if available) and the IC observations until the time of an anomaly, e.g., for a test at time 3 we have the information of only 2 RFS observations.

For a fair comparison between the two methods, we set the false alarm rate $\alpha = 1/100$ for the PC, and we equivalently set the quantile of the distribution of the ranking function under the IC process to the 0.01^{th} level, in order to derive the corresponding decision limits. As a performance metric we calculate the $F_1(t)$, given by:

$$F_1(t) = 2 \cdot tp_t/(2 \cdot tp_t + fp_t + f_n), \tag{23}$$

where tp_t and f_n are the percentages of true positives and false negative for an OOC sequence, respectively, and fp_t is the percentage of false positives (or false alarms) for an IC sequence at each time $t \in \{2, \ldots, 30\}$. Note, that we do not provide any test for $t = 1$, as PC starts testing from $t = 2$, "sacrificing" the first observation to obtain the posterior distribution. As we observe in Figure 3, the learning process of PC improves the detection performance, especially for the non-informative settings, while the prior information is beneficial, especially when the horizon of the observed sets is short. Regarding the values of the discounting factor, $\alpha_0 = 1$ achieves steadily higher performance as it uses all the information of the observed RFSs. However, other choices offer comparable performance and the flexibility for fine-tuned adaptation. Comparing the two methods, we observe that PC outperforms RF, achieving greater F_1 for all the scenarios. Even in scenarios where the RF has good

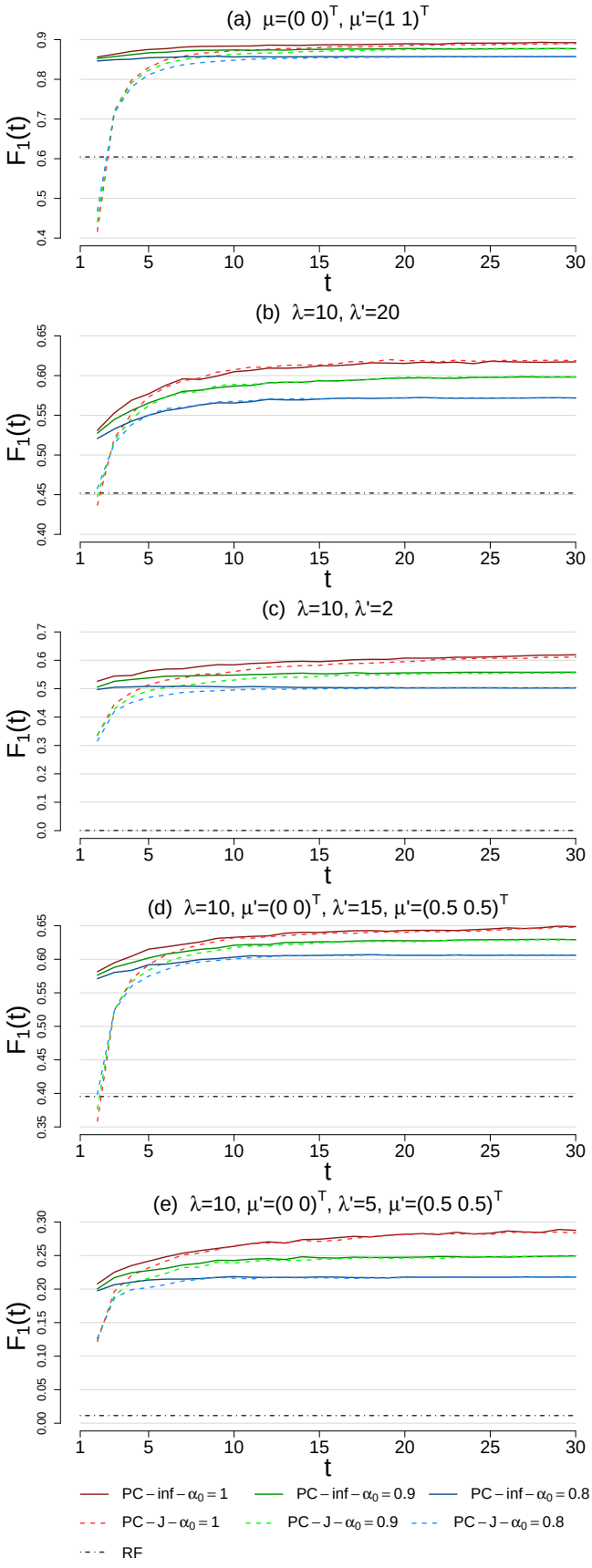

Fig. 3. The $F_1(t)$ scores for $t \in \{2, \ldots, 30\}$ of the proposed predictive checks (PC) and ranking function (RF) for 5 scenarios.

detection performance (e.g. scenario 1 or 2), the PC, even with a non-informative prior, needs only 2-3 IC observations to achieve better performance, where the parameters are assumed known. When the cardinality rate decreases, RF appears incapable of detecting anomalies. The ranking function, a weighted product of the cardinality distribution and feature likelihood, determines alarm thresholds. We raise an alarm when the ranking function's score falls below a threshold derived from the IC ranking function's quantile. For this reason, when the cardinality decreases, the likelihood factors will also decrease. Consequently, their product may not reach extremely low values, even with changes to IC parameters. This makes successful anomaly detection very difficult. On the other hand, PC is still robust in these scenarios, due to its axiomatic framework for anomaly detection.

VI. Conclusion

In this work, we developed a methodological framework to efficiently detect anomalies online for the Poisson point-patterns process while relaxing the assumption of known parameters. We deviated from the conventional approach of separating the training and testing phases; instead, we proposed a Bayesian self-starting process that sequentially estimates the unknown parameters while assessing the conformity of new sets. Additionally, we introduced a new class of power discounting posterior distributions alongside posterior predictive checks, enabling adaptive learning and robustness in detecting anomalous observations.

References

[1] V. Kumar, "Parallel and distributed computing for cybersecurity," *IEEE Distributed Systems Online*, vol. 6, no. 10, 2005.

[2] C. Spence, L. Parra, and P. Sajda, "Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model," in *Proceedings IEEE workshop on mathematical methods in biomedical image analysis (MMBIA 2001)*, pp. 3–10, IEEE, 2001.

[3] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statistical science*, vol. 17, no. 3, pp. 235–255, 2002.

[4] P. J. Diggle, *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press, 2013.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[6] D. Samariya and A. Thakkar, "A comprehensive survey of anomaly detection algorithms," *Annals of Data Science*, vol. 10, no. 3, pp. 829–850, 2023.

[7] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM computing surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[8] R. P. Mahler, *Advances in statistical multisource-multitarget information fusion*. Artech House, 2014.

[9] R. Mahler, *Statistical multisource-multitarget information fusion*. Artech, 2007.

[10] B.-T. Vo, B.-N. Vo, and A. Cantoni, "Bayesian filtering with random finite set observations," *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1313–1326, 2008.

[11] W. Yi and L. Chai, "Heterogeneous multi-sensor fusion with random finite set multi-object densities," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3399–3414, 2021.

[12] S. Papaioannou, P. Kolios, T. Theocharides, C. G. Panayiotou, and M. M. Polycarpou, "Jointly-optimized searching and tracking with random finite sets," *IEEE Transactions on Mobile Computing*, vol. 19, no. 10, pp. 2374–2391, 2019.

[13] B.-N. Vo, W.-K. Ma, and S. Singh, "Localizing an unknown time-varying number of speakers: a bayesian random finite set approach," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 4, pp. iv/1073–iv/1076 Vol. 4, 2005.

[14] W. Li, Y. Jia, J. Du, and F. Yu, "Gaussian mixture phd filter for multiple maneuvering extended targets tracking," in *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pp. 2410–2415, 2011.

[15] S. Papaioannou, P. Kolios, T. Theocharides, C. G. Panayiotou, and M. M. Polycarpou, "Probabilistic search and track with multiple mobile agents," in *2019 International Conference on Unmanned Aircraft Systems (ICUAS)*, pp. 253–262, IEEE, 2019.

[16] B. Doerr and R. Linares, "Control of large swarms via random finite set theory," in *2018 Annual American Control Conference (ACC)*, pp. 2904–2909, 2018.

[17] S. Papaioannou, P. Kolios, T. Theocharides, C. G. Panayiotou, and M. M. Polycarpou, "Decentralized search and track with multiple autonomous agents," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 909–915, 2019.

[18] S. Papaioannou, P. Kolios, T. Theocharides, C. G. Panayiotou, and M. M. Polycarpou, "A cooperative multiagent probabilistic framework for search and track missions," *IEEE Transactions on Control of Network Systems*, vol. 8, no. 2, pp. 847–858, 2020.

[19] C. Yang, L. Feng, Z. Shi, R. Lu, and K.-K. R. Choo, "A crowdsensing-based cyber-physical system for drone surveillance using random finite set theory," *ACM Trans. Cyber-Phys. Syst.*, vol. 3, aug 2019.

[20] S. Papaioannou, P. Kolios, and G. Ellinas, "Downing a rogue drone with a team of aerial radio signal jammers," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2555–2562, 2021.

[21] S. Papaioannou, P. Kolios, C. G. Panayiotou, and M. M. Polycarpou, "Cooperative simultaneous tracking and jamming for disabling a rogue drone," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7919–7926, IEEE, 2020.

[22] B.-N. Vo, N. Dam, D. Phung, Q. N. Tran, and B.-T. Vo, "Model-based learning for point pattern data," *Pattern Recognition*, vol. 84, pp. 136–151, 2018.

[23] A. M. Kamoona, A. K. Gostar, R. Tennakoon, A. Bab-Hadiashar, D. Accadia, J. Thorpe, and R. Hoseinnezhad, "Random finite set-based anomaly detection for safety monitoring in construction sites," *Ieee Access*, vol. 7, pp. 105710–105720, 2019.

[24] A. M. Kamoona, A. K. Gostar, A. Bab-Hadiashar, and R. Hoseinnezhad, "Point pattern feature-based anomaly detection for manufacturing defects, in the random finite set framework," *IEEE Access*, vol. 9, pp. 158672–158681, 2021.

[25] A. M. Kamoona, A. K. Gostar, X. Wang, M. Easton, A. Bab-Hadiashar, and R. Hoseinnezhad, "Anomaly detection of defect using energy of point pattern features within random finite set framework," *Engineering Applications of Artificial Intelligence*, vol. 130, p. 107706, 2024.

[26] P. Qiu, *Introduction to statistical process control*. CRC press, 2013.

[27] R. L. Streit and R. L. Streit, *The poisson point process*. Springer, 2010.

[28] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.

[29] X.-L. Meng, "Posterior predictive p-values," *The annals of statistics*, vol. 22, no. 3, pp. 1142–1160, 1994.

[30] R. A. Fisher, "Statistical methods for research workers," in *Breakthroughs in statistics: Methodology and distribution*, pp. 66–70, Springer, 1970.

[31] J. G. Ibrahim and M.-H. Chen, "Power prior distributions for regression models," *Statistical Science*, pp. 46–60, 2000.

[32] M. West, "Bayesian model monitoring," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 48, no. 1, pp. 70–78, 1986.

[33] H. Jeffreys, *The theory of probability*. OuP Oxford, 1998.

[34] B. P. Carlin and T. A. Louis, *Bayesian methods for data analysis*. CRC press, 2008.

[35] K. Bourazas, D. Kiagias, and P. Tsiamyrtzis, "Predictive Control Charts (PCC): A Bayesian approach in online monitoring of short runs," *Journal of Quality Technology*, vol. 54, no. 4, pp. 367–391, 2022.

[36] H. Hotelling, "Multivariate quality control-illustrated by the air testing of sample bombsights," *Techniques of statistical analysis*, 1947.

[37] C. Stein, "The admissibility of hotelling's t^2-test," *The Annals of Mathematical Statistics*, pp. 616–623, 1956.

[38] J. Prins and D. Mader, "Multivariate control charts for grouped and individual observations," *Quality Engineering*, vol. 10, no. 1, pp. 49–57, 1997.