
Article:
Sarfraz, A., Birnbaum, A., Dolan, F. et al. (3 more authors) (2025) A robust unsupervised method for outlier set detection. Knowledge-Based Systems. 114274. ISSN: 0950-7051

https://doi.org/10.1016/j.knosys.2025.114274

Author-produced version of the paper published in
Knowledge-based Systems.
Accepted on August 12, 2025.

A Robust Unsupervised Method For Outlier Set Detection

Amal Sarfraz[1,5,6,*], Abigail Birnbaum[2], Flannery Dolan[3],
Jonathan Lamontagne[2], Lyudmila Mihaylova[4] and Charles Rougé[1]

August 12, 2025

Abstract

This paper proposes a robust method that identifies sets of points that collectively deviate from typical patterns in a dataset, which it calls "outlier sets", while excluding individual points from detection. This new methodology, Outlier Set Two-step Identification (OSTI) employs a two-step approach to detect and label these outlier sets. First, it uses Gaussian Mixture Models for probabilistic clustering, identifying candidate outlier sets based on cluster weights below a hyper-parameter threshold. Second, OSTI measures the Inter-cluster Mahalanobis distance between each candidate outlier set's centroid and the overall dataset mean. OSTI then tests the null hypothesis that this distance does not significantly differ from its theoretical chi-square distribution, enabling the formal detection of outlier sets. We test OSTI systematically on 8,000 synthetic 2D datasets across various inlier configurations and thousands of possible outlier set characteristics. Results show OSTI robustly and consistently detects outlier sets with an average F1 score of 0.92 and an average purity (the degree to which outlier sets identified correspond to those generated synthetically, i.e., our ground truth) of 98.58%. We also compare OSTI with state-of-the-art outlier detection methods, to illuminate how OSTI fills a gap as a tool for the exclusive detection of outlier sets.

Keywords: Outlier sets, Outlier Set Two-step Identification (OSTI), Gaussian mixture models, Inter-cluster Mahalanobis distance

Key points

- New approach to label clustered sets as outliers sets instead of individual points.
- Our two-step method combines clustering and distance-based outlier set detection.
- Traditional outlier detection methods are not designed to tackle this problem.
- Our method (OSTI) reliably identifies outlier sets across 8,000 synthetic datasets.

[1]School of Mechanical, Aerospace and Civil Engineering, University of Sheffield, United Kingdom
[2]Department of Civil and Environmental Engineering, Tufts University, United States
[3]RAND Corporation, California, United States
[4]School of Electrical and Electronic Engineering, University of Sheffield, United Kingdom
[5]Institute of Environmental Sciences and Engineering, National University of Sciences and Technology, Pakistan
[6]Department of Physical Geography, Utrecht University, The Netherlands
*Corresponding author

1 Introduction

Outliers are commonly defined as data points deviating significantly from the remainder of dataset (*Grubbs*, 1969; *Grubbs and Beck*, 1972; *Barnett*, 1978; *Hawkins*, 1980; *Beckman and Cook*, 1983). This definition with emphasis on individual observations has endured different conceptualizations of outliers and methods to detect them. For instance, outliers can deviate significantly from the remainder of the dataset either in absolute terms – they are *global outliers* – or within a specific context – they are *contextual outliers*; or they can be anomalous because they are part of a group of points that collectively stand out – they are *clustered or collective outliers* (*Divya and Babu*, 2016; *Singh and Upadhyaya*, 2012; *Zimek et al.*, 2014). In all these cases, including *clustered or collective outliers*, the aim is to label each observation as outlier (or inlier) individually, rather than characterizing the cluster or collective that they form (*Gao et al.*, 2022; *Li et al.*, 2022).

However, there is a growing recognition that outliers manifest as coordinated phenomena where the collective behavior of seemingly typical individual points reveals systematic deviations, coordinated threats, or emergent patterns that cannot be captured through individual point analysis (*Yu et al.*, 2015; *Thudumu et al.*, 2020). This limitation becomes particularly problematic in complex, interconnected systems where the most critical insights emerge from understanding groups of related observations as single analytical units rather than as collections of individually labeled points. This points to a qualitative difference between identifying and interpreting points individually or collectively. Indeed, while individual outliers typically indicate isolated anomalies or measurement errors, outlier sets expose coordinated threats, systematic irregularities, and collective behaviors that emerge from the interactions between multiple data points - patterns that cannot be understood by analyzing individual observations in isolation (*Saha et al.*, 2018; *Breunig et al.*, 2000; *Bai et al.*, 2016). To formalize this qualitative difference, we introduce the concept of "outlier sets" - cohesive groups of data points that collectively deviate from expected patterns, and which we are interested in detecting and analyzing as unified entities rather than individual outlier observations.

Formalizing outlier sets as a type of object separate from point outliers would be valuable in several fields where collective outliers provide actionable insights unattainable through individual point detection. In big data analytics, processing vast information streams, sets of outliers reveal systematic biases and collective behavioral patterns that detection methods for individual outliers are not meant to find, particularly in high-dimensional datasets where many approaches suffer from the curse of dimensionality (*Thudumu et al.*, 2020). Digital twin-based outlier detection frameworks highlight that identifying collective outlier patterns across data streams is critical for capturing underlying distribution shifts (*Gupta et al.*, 2024).

Other current efforts looking at large future uncertainties - climatic, socio-economic and / or technological, to name a few - lead to the generation of thousands to millions of scenarios to understand the behavior of complex systems across these large uncertainty spaces an approach often referred to as exploratory modeling (*Kwakkel and Pruyt*, 2013; *Moallemi et al.*, 2020). In such large ensembles, identifying individual outliers is often not of interest. Instead, more meaningful insights are generally obtained from identifying groups of scenarios that collectively represent meaningful deviations from typical patterns, and from understanding what these scenarios have in common. For instance, datasets obtained by simulating large scenario ensembles are increasingly used to understand the fac-

tors that lead to remarkable outcomes in climate change, climate policy and associated socio-economic scenarios (*Eyring et al.*, 2016; *Lamontagne et al.*, 2018; *Dolan et al.*, 2021; *Byers et al.*, 2022; *Dekker et al.*, 2023). Transportation infrastructure requires collective monitoring where cascading failures begin as coordinated outlier sets across interconnected systems, and early detection of these collective patterns prevents widespread service disruptions (*Basak et al.*, 2019). In cybersecurity, Advanced Persistent Threats operate through coordinated actions where individual network events appear normal but collectively reveal sophisticated attack patterns that traditional individual-based intrusion detection systems routinely miss, enabling prolonged unauthorized access (*Benabderrahmane et al.*, 2024). Financial systems have evolved to recognize that sophisticated fraud schemes like coordinated market manipulation and money laundering manifest as collective transaction patterns rather than individual suspicious activities, requiring detection of sets of outliers (*Mazzarisi et al.*, 2024). Similarly, in meteorology, several recent studies point to the need to characterize emerging extreme weather events as spatio-temporal anomalous patterns, rather than through large collections of individual grid points involved over the event's time-span (*Ramachandra et al.*, 2016; *Barz et al.*, 2017; *Coelho et al.*, 2008). Beyond these examples, interest in detecting and describing sets of outliers and describing their anomalous behaviors, not as individual points, but as spatial and/or temporal clusters has arisen in fields as varied as finance (*Lee et al.*, 2022), healthcare (*Allenby et al.*, 2021) and smart grid analytics (*Madabhushi and Dewri*, 2023).

To address this emerging need for collective identification of sets of clustered outliers, the main contributions of this paper are summarized as follows:

- We formalize the outlier set detection problem and demonstrate that existing state-of-the-art methods cannot solve it, establishing a clear methodological gap where current approaches detect individual outliers but fail to identify coherent anomalous groups as unified entities.

- We introduce a novel two-step methodology for the exclusive detection of outlier sets, and call it Outlier Set Two-step Identification (OSTI). It identifies potential outlier sets with probabilistic clustering, and combines it with Inter-cluster Mahalanobis distance measurement and rigorous hypothesis testing to formally detect a cluster as an outlier set.

- We demonstrate through an example dataset that our methodology, OSTI, is qualitatively different from existing outlier detection methods in that no other methods can detect outlier sets while excluding individual points.

- We conduct a systematic evaluation of the effectiveness of OSTI in two dimensions, through the generation of thousands of synthetic datasets varying inlier and outlier configurations. It demonstrates the effectiveness of our approach in correctly identifying outlier sets with a high degree of purity i.e., the detected outlier sets coincide with the synthetically generated outlier sets with a high degree of accuracy. We also provide evidence that results are robust to hyperparameter choices, making OSTI convenient to implement.

The rest of the paper is structured as follows. We analyze the state-of-the-art outlier detection methods for outlier set detection (Section 2). Then, we introduce the OSTI methodology (Section 3). Next, we describe the experimental setup, including the large-scale synthetic dataset generation process

used to evaluate OSTI (Section 4). We then present results (Section 5), followed by a discussion of key findings (Section 6), and conclude with a summary of contributions and future directions (Section 7).

2 Related Work

Outlier detection methods can be broadly classified into distance-based, density-based, clustering-based and ensemble-based methods (*Campos et al.*, 2016a; *Aggarwal and Sathe*, 2015; *Mandhare and Idate*, 2017; *Hodge and Austin*, 2004) to detect the different types of (individual or point) outliers mentioned in the introductory paragraph of Section 1.

Distance-based methods assess the outlier status of a point based on its distance from its neighbors (*Kriegel et al.*, 2008; *Ghorbani*, 2019; *Maesschalck et al.*, 2000; *Ramaswamy et al.*, 2000). They are well-suited to the detection of point and contextual outliers by identifying points far from others. For instance, Mahalanobis distance (*Mahalanobis*, 1933) measures the separation between a point and a distribution while removing cross-variable correlation via an orthogonal coordinate transformation. This makes it effective for detecting outliers that appear normal when variables are considered independently in highly correlated multivariate datasets (*Li et al.*, 2019). However, it is not set up to recognize structured deviations that form coherent groups of points, i.e., outlier sets.

Density-based methods identify regions in the data space where the density of points is significantly lower than in the surrounding areas. Points in these low-density regions are considered outliers. Density-based methods are suited for identifying point and contextual outliers in sparse areas; however, they can typically struggle in datasets where inliers are in regions of varying densities (*Breunig et al.*, 2000; *Bai et al.*, 2016). For this reason, more recent methods such as the Relative-KNN-kernel density-based clustering algorithm (REDCLAN) (*Saha et al.*, 2018) operates on the principle of relative density rather than absolute density to detect clusters. It includes a weighted rank-based anomaly detection component that helps detect outliers relative to the identified clusters. REDCLAN effectively groups similar density points into clusters while identifying those that do not conform to the overall data distribution as outliers. It is unclear how density-based methods would handle outlier sets.

Clustering-based methods operate by organizing data points into clusters and can detect various types of outliers through two broad approaches. In the first approach, outliers are characterized as data points that either do not belong to any cluster or are significantly distant from the nearest cluster (*Ester et al.*, 1996). In the second approach, outliers are identified as points that form micro-clusters (*Breunig et al.*, 2000) - small groups of data points that are markedly different from the main clusters in the dataset. Several clustering-based outlier detection methods illustrate these approaches. For instance, the Cluster-Based Local Outlier Factor (CBLOF) (*He et al.*, 2003) detects outliers by clustering the dataset into small and large clusters based on initialization parameters α and β, which act as thresholds and weighting factors respectively. Data points in small clusters located near large clusters are identified as outliers, enabling the identification of small groups of isolated points. However, CBLOF's effectiveness depends heavily on initialization parameters and lacks verification mechanisms for identified outlier clusters. Another method, Outlier Detection with Explicit Micro-Cluster Assignments (D.MCA) (*Jiang et al.*, 2022), blends advanced sampling strategies, pruning, and iterative refinement to detect outliers while assigning them to explicit micro-clusters. Despite its sophisticated approach,

D.MCA's point-wise scoring can fragment outlier identification, potentially missing some points within outlier clusters. The relationship between clustering and outlier detection extends beyond using clustering for outlier identification. Jiang et al. (*Jiang et al.*, 2016) demonstrated that outlier detection techniques can enhance clustering performance through improved initialization strategies. They proposed initialization algorithms for K-modes clustering that incorporate outlier detection techniques, establishing a bidirectional relationship where clustering aids outlier detection and outlier detection improves clustering initialization. Their work reinforces the principle that outliers should not be selected as initial cluster centers, as this can lead to suboptimal clustering results. This bidirectional relationship highlights a fundamental limitation in existing approaches: most methods focus on individual point detection rather than considering the collective behavior of sets of outliers. This observation reinforces our motivation for developing methods that can detect entire outlier sets as cohesive entities, rather than fragmenting them into individual outlier points. Despite their demonstrated effectiveness, clustering-based outlier detection methods (*Sánchez Vinces et al.*, 2025) remain limited by their focus on individual data points, overlooking cohesive outlier groups and thus constraining their utility for detecting collective anomalies.

Ensemble-based methods employ multiple algorithms to identify outliers in a given dataset for increased accuracy and adaptability (*Sahu et al.*, 2021; *Li and Zhang*, 2023; *Ouyang et al.*, 2021). Ensemble-based methods are versatile in detecting point, collective or clustered and contextual outliers by leveraging multiple detection strategies. For instance, Feature Bagging operates by combining multiple outlier detection algorithms over different feature subsets (*Lazarevic and Kumar*, 2005). Used in conjunction with Isolation Forest (*Liu et al.*, 2008), each algorithm produces an outlier score for each data point, and these scores are combined to detect higher-quality outliers. This approach is particularly effective in high-dimensional and noisy datasets but is aimed at identifying outliers along dataset boundaries rather than coherent groups. A recent unsupervised outlier detection method, Differential Potential Spread Loss (DPSL) (*Gao et al.*, 2022) detects global, local, and clustered outliers simultaneously using potential chains based on nearest neighbor relationships. It constructs hierarchical potential peak points and measures anomaly degrees through distance ratios and isolation radius. DPSL requires careful parameter tuning and has limited scalability for large datasets due to point-by-point evaluation.

Even though some clustering and ensemble-based methods can detect small groups of points, none of these methods aim at exclusively finding sets of clustered outliers - that is, excluding isolated outlying points. Yet, there is a growing need to detect collective outlying or anomalous behaviors in data (*Feroze et al.*, 2021). Several outlier detection methods do not fit neatly into traditional categories mentioned above yet remain significant. These methods employ alternative strategies such as boundary construction in feature spaces, statistical distribution modeling, and dimensionality reduction techniques to identify outliers. For instance, Deep Support Vector Data Description (Deep SVDD) (*Ruff et al.*, 2018) employs a neural network trained to map input data into a hypersphere of minimum volume in the learned feature space, directly optimizing an anomaly detection objective. Deep SVDD is particularly effective for high-dimensional data where traditional kernel methods fail due to computational constraints, as it scales linearly with dataset size and requires no data storage for prediction. In contrast to Deep SVDD's neural network-based feature mapping, Copula-based Outlier

Detection (COPOD) (*Li et al.*, 2020) leverages copula functions to model the dependence structure between features, enabling the detection of outliers based on deviations from expected joint distributions. COPOD is particularly powerful for multivariate outlier detection as it can capture complex dependencies between variables. COPOD computes outlier scores by measuring how well each data point fits the learned copula-based dependency model, making it effective for detecting contextual outliers. However, COPOD's point-wise evaluation approach prevents it from recognizing entire clusters as cohesive outlier sets, instead treating each anomalous point as an isolated deviation from the modeled dependencies.

Complementing these approaches, Lightweight On-line Detector of Anomalies (LODA) (*Pevnỳ*, 2016) utilizes sparse random projections to transform high-dimensional data into multiple one-dimensional spaces, where outlier detection is performed using histogram-based density estimation. LODA is computationally efficient and particularly effective for high-dimensional datasets, as it reduces the curse of dimensionality by working with multiple random projections simultaneously. The method aggregates outlier scores across all projections to provide a final anomaly score for each data point, making it robust against noise and capable of detecting various types of outliers. However, LODA's projection based strategy suffers from the same fundamental limitation as the previous methods: it fragments outlier set detection by evaluating points individually across different projection spaces.

Despite their diverse approaches, these methods share a common limitation when applied to outlier set detection: they are designed to evaluate individual data points rather than recognizing cohesive points as outlier sets. This limitation is challenging for large scenario ensembles, exploratory modeling, and applications requiring understanding of outlying sets patterns. The increasing complexity of datasets emphasizes the need for methods that efficiently identify and characterize meaningful outlier sets patterns involving groups of related observations, which existing approaches fail to address effectively.

3 Methodology

This section describes our proposed methodology Outlier Set Two-step Identification (OSTI). OSTI integrates probabilistic clustering and a Mahalanobis distance-based statistical test to detect data points that cluster together as outlier sets. Our approach involves two main steps, (A) the identification of candidate outlier sets with Gaussian Mixture Models, and (B) the identification of outlier sets with a chi-square test based on Inter-cluster Mahalanobis distance.

A. **Identification of candidate outlier sets using Gaussian Mixture Models**

We perform probabilistic clustering, using Gaussian Mixture Models (GMM) due to their well-recognized capacity to handle clusters of different sizes, shapes, and densities (*Day*, 1969). GMMs assume all the data points are generated from a mixture of a finite number of Gaussian distributions. In this study, we emphasize the utility of GMMs for extracting component weights and highlight the relationship between number of clusters and component weights. The probability

density function of a GMM is given by equation (1):

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{1}$$

where K is the number of mixtures (or clusters), and the k-th mixture component has a Gaussian distribution $\mathcal{N}(x|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$, and the π_k are called the mixing weights. They represent the prior probabilities of different clusters or components, and sum up to 1. Each weight π_k reflects the proportion of the data that belongs to that cluster, i.e., a higher π_k means that the k^{th} cluster has more points.

Next, we label candidate outlier sets as the components of the GMM that are below a weight threshold π_{th}. The rationale for this is that above that threshold, the presence of that cluster is a feature of the data rather than a potential outlier set. Setting a threshold is also key to avoiding the detection of large group of inliers as a false positive in Step B. On the other end of the spectrum, setting the weight threshold too low risks missing relatively large outlier sets.

GMMs are sensitive to the choice of the number of clusters K, a hyperparameter that needs to be decided in advance (*Kasture and Gadge*, 2012), and K directly determines the weight of typical outlier sets. Therefore the choice of K and π_{th} are linked, as the choice of K must not impede the identification of outlier clusters. Multiple methods can be used for choosing K, e.g., Bayesian Information Criterion or Akaike Information Criterion (*Cheong and Lee*, 2008; *Schwarz*, 1978; *Ishioka et al.*, 2005). Our practice, based on several datasets similar to Figure 4, suggests that it is appropriate to choose K such that $1/K$ is larger than, but close to, the threshold π_{th}, so the threshold applies to smaller-than-average clusters.

B. **Outlier set identification with a chi-square test based on Inter-cluster Mahalanobis distance**

Mahalanobis distance is a measure of the distance between a point and a distribution that accounts for its covariance structure (*Mahalanobis*, 1933), as defined by the following equation (2).

$$\mathrm{MD}(\mathbf{x}; \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}), \tag{2}$$

where MD is the squared Mahalanobis distance, computed between a point represented by vector $\mathbf{x}$ in N-dimensional space, and vector $\boldsymbol{\mu}$ is the distribution mean. $\boldsymbol{\Sigma}^{-1}$ is the inverse covariance matrix used to transform into an orthogonal set of coordinates where each variable is independent, and T denotes the transpose operation.

Instead of measuring the deviation between individual points and a distribution's center, in OSTI we compute the distance between the mean of a candidate outlier cluster and mean of the rest of the dataset and refer to it as the Inter-cluster Mahalanobis distance (IMD).

We define IMD in equation (3) by updating equation (2) as follows:

$$IMD(\boldsymbol{\mu}_{\mathrm{cl}}; \boldsymbol{\mu}) = (\boldsymbol{\mu}_{\mathrm{cl}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{\mathrm{cl}} - \boldsymbol{\mu}), \tag{3}$$

where $\boldsymbol{\mu}_{\text{cl}}$ represents the mean of the cluster, and $\boldsymbol{\mu}$ is the mean vector of the entire dataset. IMD accounts for the variance and covariance of the data in exactly the same way as MD, and aids in identifying clusters that are farthest from the mean of the dataset. By utilizing IMD rather than Euclidean distance, we ensure that scale differences across dimensions have no impact on the identification of outlier sets.

Next, we test for each identified candidate outlier and set the null Hypothesis (H_0) that its mean $\boldsymbol{\mu}_{\text{cl}}$ is the same as that of the rest of the dataset, $\boldsymbol{\mu}$. The alternative hypothesis (H_1) is that $\boldsymbol{\mu}_{\text{cl}}$ is significantly different from $\boldsymbol{\mu}$. Under H_0, IMD, as defined in equation (3) as the sum of the squares of independent standard normal variables, follows by definition a chi-square distribution with N degrees of freedom (*Ghorbani*, 2019). Therefore, we reject H_0 at the significance level α when (and only when) $\boldsymbol{\mu}_{\text{cl}}$ is greater than the $(1-\alpha/2)$ quantile of the chi-square distribution with N degrees of freedom $\chi^2(N)$ as represented in equation 4:

$$p_{\text{value}} = P\left(IMD(\boldsymbol{\mu}_{\text{cl}}; \boldsymbol{\mu}) > \chi^2(N)\right) \leq \alpha, \tag{4}$$

The pseudo code for the implementation of OSTI is given in Algorithm 1.

Algorithm 1 Outlier Set Two-step Identification (OSTI)

Require: Dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$, number of clusters K, weight threshold π_{th}, significance level α
Ensure: Set of detected outlier sets $\mathcal{O}$

Step A: GMM Clustering for Candidate Identification
Initialize GMM with K components
Fit GMM on dataset $\mathbf{X}$ using EM algorithm
Extract cluster parameters: $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ for $k = 1, \ldots, K$
$\mathcal{C}_{candidates} \leftarrow \emptyset$
for $k = 1$ to K **do**
  if $\pi_k \leq \pi_{th}$ **then**
    $\mathcal{C}_{candidates} \leftarrow \mathcal{C}_{candidates} \cup \{\text{cluster } k\}$
  end if
end for
Step B: Statistical Verification using Chi-square Test
$\mathcal{O} \leftarrow \emptyset$
Compute dataset mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
for each cluster c in $\mathcal{C}_{candidates}$ **do**
  Compute cluster centroid $\boldsymbol{\mu}_{cl}$
  Calculate $IMD = (\boldsymbol{\mu}_{cl} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{cl} - \boldsymbol{\mu})$
  Compute p-value $= P(\chi^2(d) > IMD)$
  if p-value $\leq \alpha$ **then**
    $\mathcal{O} \leftarrow \mathcal{O} \cup \{\text{cluster } c\}$
  end if
end for
return $\mathcal{O}$

4 Evaluation of OSTI

This paper formally poses the new problem of detecting outlier sets, to the exclusion of individual outliers. It then proposes a new methodology, OSTI, to solve this problem. We have two main evaluation objectives. Our first objective is to verify that OSTI performs qualitatively differently from existing individual outlier detection methods. Our second objective is to verify across a large sample of diverse inlier / outlier set configurations (1) that OSTI's first clustering step correctly identifies outlier sets, and (2) that the statistical test in OSTI's second step is powerful for the detection of outlier sets.

For these tasks, we cannot rely on existing benchmark repositories such as the Outlier Detection DataSets (ODDS) Library (e.g., (*Campos et al.*, 2016b; *Rayana*, 2016)) and UCI Machine Learning Repository (*Kelly et al.*, 2017). Indeed, these and similar benchmarks datasets label each observation as either an outlier or inlier, but they do not label coherent groups of outliers that should be treated as unified entities (*Campos et al.*, 2016c; *Emmott et al.*, 2013). They are not benchmarks for the outlier set detection problem that this paper addresses - which is expected, as this is a formally new problem. OSTI solves a problem that is qualitatively different from the problem of correctly labelling individual observations because it labels whole clusters as inliers or outlier sets. In other words, OSTI is explicitly not suited to label single observations, making it inappropriate to apply to existing benchmarks.

However, we need to check whether conversely, methods designed for labelling individual points as inliers and outliers, are appropriate or not for the problem of exclusive detection of outlier sets. This corresponds to our first evaluation objective, and for this we use an existing real-world dataset for which we know the ground truth, as described in detail in Section 4.1.

After this, we use synthetic dataset to provide a clearly interpretable evaluation of OSTI's capabilities – our second evaluation objective. This is because synthetic datasets enable meaningful evaluation with clear anomaly characteristics compared to benchmark datasets with unknown collective anomaly properties (*Steinbuss and Böhm*, 2021); they also allow controlled experimentation in a way real-world benchmarks do not (*Emmott et al.*, 2015). We detail our synthetic dataset approach in Section 4.2.

4.1 Cross-Method Comparison

The cross-method comparison serves primarily as a qualitative demonstration of the fundamental methodological differences between OSTI and existing approaches rather than comprehensive quantitative benchmarking. We utilize a single real-world dataset for this purpose because existing benchmark repositories lack established ground truth for outlier sets, making quantitative comparison across multiple real-world datasets methodologically challenging. The IRB dataset provides an ideal proof-of-concept case where distinct outlier sets are visually apparent, enabling clear demonstration of how existing methods either fragment outlier set detection or incorrectly flag scattered individual points. This qualitative analysis effectively illustrates the conceptual gap that OSTI addresses, while our comprehensive quantitative evaluation is conducted using synthetic datasets with known ground truth in subsequent sections.

First, we evaluate OSTI against other state-of-the-art methods using a subset from a large ensemble of scenarios where ground truth is known to detect outlier sets. An example of outlying behavior is represented in Figure 1, which shows irrigation-scarcity relationships for 3,000 scenarios of the Indus

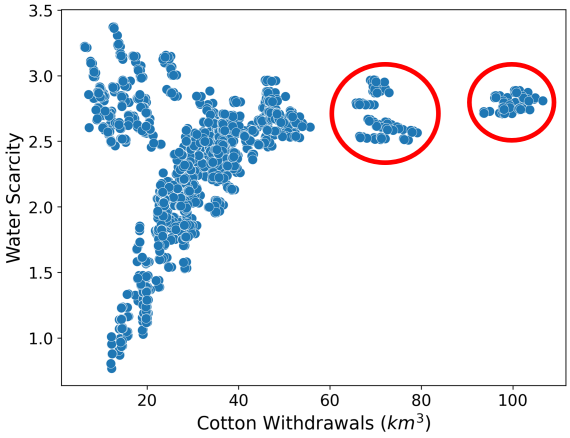

Figure 1: Cotton irrigation water withdrawals (km^3) versus water scarcity for the Indus River Basin in 2100. Scatter plot showing 3,000 scenarios with two distinct outlier sets circled in red, each representing a cluster of points that collectively deviates from the main distribution pattern, indicating unusual combinations of irrigation withdrawals and water scarcity levels.

River Basin (IRB) in 2100. These scenarios are extracted from a 6TB dataset generated using the Global Change Analysis Model (GCAM) (*Dolan et al.*, 2021). GCAM has played an integral role in Intergovernmental Panel on Climate Change (IPCC) reports for over two decades, offering in-depth insights into socio-economics, water supply-demand dynamics, and land use changes (*Calvin et al.*, 2017; *Ou et al.*, 2021), and enabling a comprehensive exploration of interactions between human and Earth system processes (*Calvin et al.*, 2019). Figure 1 clearly shows two sets of outliers, which we circled in red. We use this IRB dataset to implement multiple approaches discussed in Section 2 with different properties to evaluate their respective ability to detect these outlier sets.

To systematically evaluate existing methods, we implemented representative approaches from four major categories of outlier detection. For distance-based detection, Mahalanobis distance was implemented. REDCLAN was employed as a representative of density-based methods. Clustering-based approaches included Clustering-Based Local Outlier Factor (CBLOF) and Detection with Explicit Micro-Cluster Assignments (D.MCA). We further evaluated two ensemble-based methods: Feature Bagging and Differential Potential Spread Loss (DPSL). Lastly, we tested three additional techniques: Deep Support Vector Data Description (Deep SVDD), Lightweight On-line Detector of Anomalies (LODA), and Copula-Based Outlier Detection (COPOD).

We used the Python library PyOD (*Zhao et al.*, 2019) to implement Feature Bagging with Isolation Forest and CBLOF. For Mahalanobis distance, REDCLAN, D.MCA, DPSL, SVDD, COPOD, and LODA, we implemented custom code based on the original algorithms described in their respective papers.

4.2 Synthetic Datasets

We conduct systematic validation against thousands of synthetic datasets where different ground truth conditions are defined. This allows rigorous testing of OSTI's performance across different inlier configurations and outlier set characteristics. These synthetic datasets provide an objective benchmark for assessing OSTI's outlier set detection capabilities by answering both the following questions: (Q1) Can clusters of outlying points be identified as outlier sets?, and (Q2) To what extent do identified outlier sets overlap with true outlier sets? Firstly we describe the synthetic generation of datasets (Section 4.2.1). Secondly, we establish the ground truth against which we apply OSTI (Section 4.2.2). Thirdly, we define clear performance metrics to answer both questions (Section 4.2.3). And finally we detail the numerical experiments we conduct to evaluate OSTI (Section 4.2.4).

4.2.1 Synthetic Datasets Generation

We use *make_ blobs*, a function from the Python library scikit-learn (*Pedregosa et al.*, 2011) (see Appendix C), to generate inlier sets with 1,500 data points, 15 centers, 6 standard deviation and the bounding range within which the cluster centers are randomly generated is (-10,10) for each of the 2D features, ensuring variability in the spatial distribution of the generated data points. Next, four different shape configurations are applied to these inliers, i.e., circle, ellipse, triangle and irregular with no transformation. Still using *make_ blobs*, we also generate smaller floating sets of clustered points, potentially detectable as outlier sets based on three parameters. d_{out} represents the radial distance from the centroid of the inlier set in a polar coordinate system, ranging from 0 to 100 units. ϑ_{out} denotes the angular position of the floating sets, ranging from 0 to 360 degrees. Standard deviation σ_{out}, ranging from 1 to 10 units, measures the spread of data points within the floating sets. We explore the potential of OSTI in cases where only one floating set is generated, and in cases where two are generated, as follows:

- In Case 1, we introduce a single floating set. For each inlier set configuration (circle, ellipse, triangle, irregular), and a fixed size of the floating set, we generate 1,000 synthetic datasets using Latin hypercube sampling (LHS) (*McKay et al.*, 2000) of the three parameters (d_{out}, ϑ_{out}, σ_{out}) across the ranges stated above.
- In Case 2, we introduce two floating sets to evaluate whether and how the detection of one outlier set interferes with that of the other. They have respective parameters (d_{out1},ϑ_{out1},σ_{out1}) and (d_{out2},ϑ_{out2},σ_{out2}). The centroid of the second floating set is at an angle $\vartheta_{\text{out1}} + 180 + \vartheta_{\text{out2}}$ in degrees, so that the two outlier sets can only overlap when ϑ_{out2} is close to 180 degrees. Furthermore, the likelihood of both floating sets overlapping is less than 10% due to the manner in which they are generated. For each inlier set configuration (circle, ellipse, triangle, irregular), and a fixed size of the floating sets, we generate 1,000 synthetic datasets using LHS of the six parameters across the ranges stated above.

4.2.2 Ground Truth

In each synthetically generated dataset, we need to label the floating set(s) as either an outlier set or not. In some cases, this is straightforward. For instance, the sets of red triangles in Figure 2 (a) and

(b) have their convex hulls fully separated from those of the inlier sets, and clearly are outlier sets. Conversely, still in Figure 2 (b) the floating set represented with black squares is not, and its convex hull is almost entirely inside the convex hull of the inlier set. In another case however, such as Figure 2 (c) the convex hull of the floating set is overlapping with that of the inlier set (irregular shape).

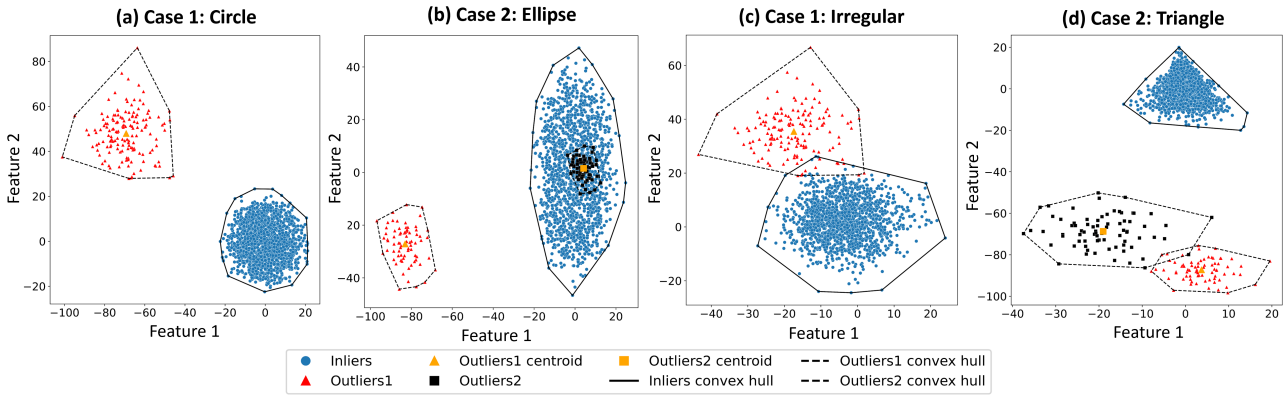

Figure 2: Illustration of ground truth Explanation for Case 1: Circle (One outlier set) (b) Case 2: Ellipse (Two outlier sets), (c) Case 1 for Irregular (no transform) and (d) Case 2 for Triangle

To account for these cases we propose two different types of ground truths. The strict ground truth labels a floating set as an outlier set if it has no point within the convex hull of the inlier set. Under the relaxed ground truth however, a floating set is labeled as an outlier set when the set's centroid lies outside the convex hull of the inlier set. These two types of ground truths agree in Figure 2 (a) and (b), where red-triangle floating sets are labeled as outlier sets whereas the black-square floating set is not. In panel Figure 2 (c), the floating set is considered an outlier set only under the relaxed ground truth, not under the strict ground truth.

Note this definition of ground truth only applies to the position of outlier sets with respect to inlier sets. For instance in Figure 2 (d), both the floating sets (in red triangles and black squares) are clearly outlying under both ground truths. They are also overlapping because they are generated with ϑ_{out2} close to 180 degrees as explained in section 4.2.1. They are considered as two separate outlier sets, as a way to understand how the methods works in such cases.

4.2.3 Performance Metrics

To evaluate the ability to detect outlier sets, for each set of 1,000 synthetic datasets we call a result positive when an outlier set is identified, and record true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). We use these to compute the classic metrics of precision P (*Powers*, 2011) and recall R (*Van Rijsbergen*, 1977) as defined in equation 5:

$$P = \frac{TP}{TP + FP} \quad \text{and} \quad R = \frac{TP}{TP + FN} \tag{5}$$

From this we derive the F1 score (*Chinchor and Sundheim*, 1993; *Sasaki et al.*, 2007) defined by equation **??**:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{6}$$

We also need to evaluate the extent to which identified outlier sets overlap with the synthetically generated set. For this we define a purity metric which is a point wise comparison of the sets. Thus, the purity P_i defined by equation 7 of the i-th cluster can be defined as the fraction of outliers within that cluster.

$$P_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \text{outlier\_labels}_j, \tag{7}$$

where outlier_labels_j is the label of the j-th point in the i-th cluster and n_i is the total number of points in the i-th cluster. This purity metric is quantified on a scale from 0% to 100%, where 0% means all points in the detected outlier sets were generated as inliers, and 100% signifies the complete point-wise identification of an outlier set. This metric evaluates the performance of clustering for the aim defined in the OSTI methodology, i.e., the identification of candidate outlier sets. For this reason, we do no implement traditional clustering metrics such as Davies-Bouldin Index (*Davies and Bouldin*, 1979) or Silhouette Index (*Rousseeuw*, 1987) which have the distinct objective of evaluating the performance of clustering throughout the whole dataset. Purity metrics are evaluated only on true positive results, as it is irrelevant to know whether a false positive set is constituted of inlier points or floating sets.

4.2.4 Design of experiments

We generated a total of 8,000 synthetic datasets, comprising 1,000 synthetic datasets for each of the four inlier shapes in Case 1 with 150 points in the floating set, and 1,000 synthetic datasets for each of the four inlier shapes in Case 2 with 75 points in each floating set. For each dataset, we applied OSTI with a threshold weight $\pi_{th} = 0.1$, ensuring no cluster representing more than 10% of the dataset can be a candidate outlier set, and used $K = 8$ clusters for Gaussian Mixture Model (GMM) clustering (see Appendix B). To benchmark our method, we applied the best-performing approach from Section 5.1 to all synthetic datasets. To check OSTI's robustness to hyperparameters, we conducted two additional experiments: first, re-applying OSTI to the same 8,000 synthetic datasets with $K = 7$ and $K = 9$, and second, generating 8,000 new synthetic datasets with only one-third the number of points in each floating set to verify detection of smaller outlier sets while maintaining the weight threshold $\pi_{th} = 0.1$. We repeated these experiments five times for each dataset to produce robust results and demonstrate the low variability between experiment sets.

All experiments were conducted on a system equipped with a 64-bit OS, 24 GB of RAM, and an Intel(R) Core(TM) i7-9700 CPU at 3.00GHz. All the code, results, and datasets (*Sarfraz et al.*, 2025) associated with this work can be found in the University of Sheffield's data repository `https://doi.org/10.15131/shef.data.28227974.v2`.

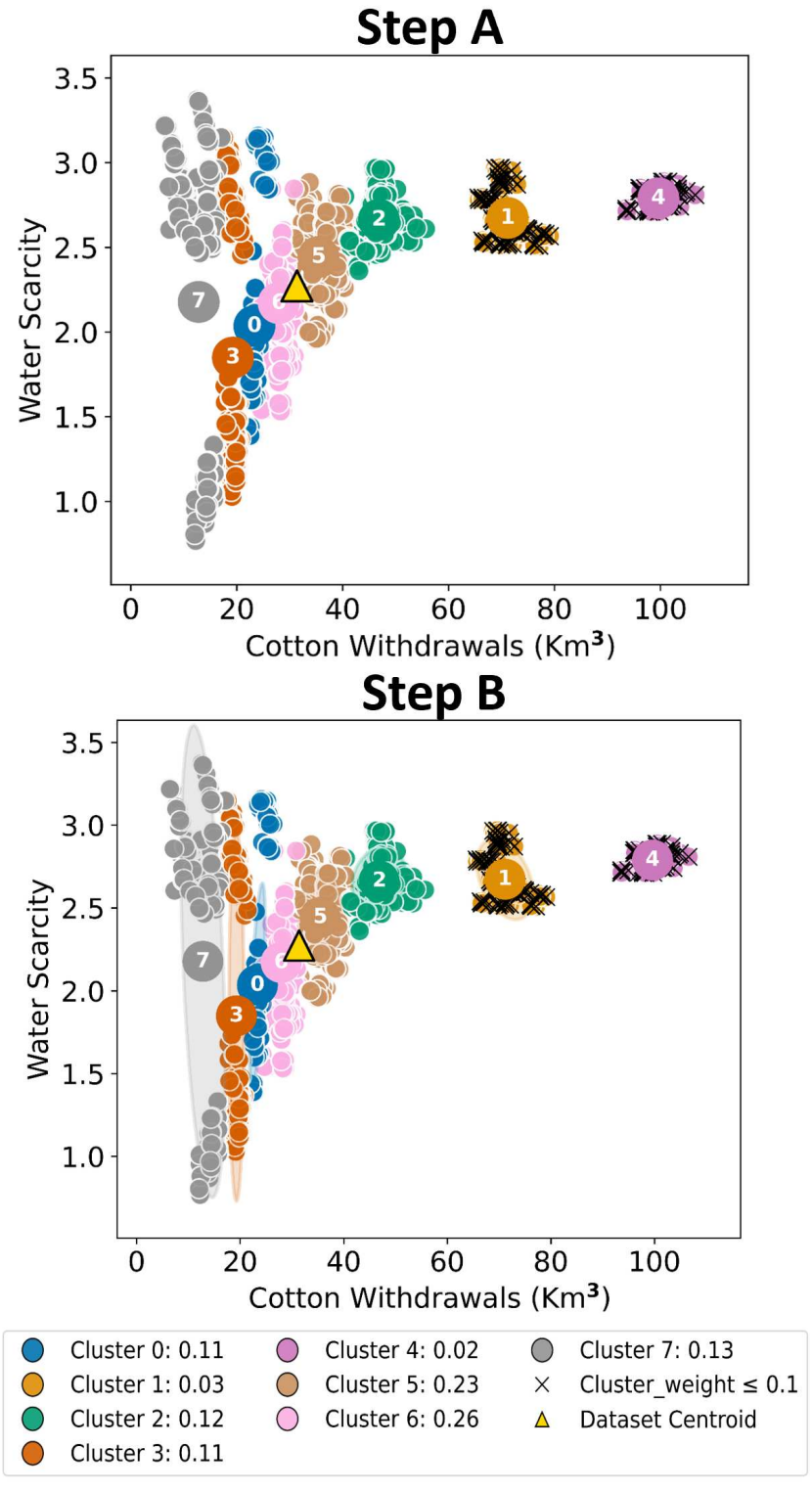

Figure 3: OSTI implementation on the dataset in Figure 1. Step A: GMM clustering with eight clusters (weights in legend); clusters with weights ≤ 0.1 are marked as candidate outliers ('x'). Step B: IMD evaluation, with shaded regions representing the covariance structure.

5 Results

The rest of this section details evaluation results of OSTI. We first assess outlier detection methods on the IRB dataset (Section 5.1). Then we evaluate OSTI's performance on synthetics datasets, through F1 score (Section 5.2) as well as purity metrics (Section 5.3), and we benchmark OSTI against existing approaches on synthetic data (Section 5.4). The analysis concludes by examining OSTI's hyperparameter robustness (Section 5.5).

5.1 Cross-Method Comparison Results

First, we demonstrate OSTI's two-step approach applied to the IRB dataset in Figure 3. In Step A, GMM clustering identifies 8 clusters (shown in different colors) with their respective weights, marked by cluster centroids. Several clusters (1, 4) have weights ≤ 0.1 (marked with 'x'), identifying them as candidate outlier sets. The dataset centroid is marked with a yellow triangle. In Step B, these candidate clusters are evaluated using IMD. The shaded regions highlight the covariance structure considered when computing distances between cluster means and the dataset centroid. This two-step process successfully identifies the two distinct outlier sets while maintaining cluster coherence.

We now examine how the outlier detection methods (detailed in Section 4.1) tackle our problem of detecting outlier sets, to the exclusion of outlying individual points. Across Figure 4, points labeled as outliers are represented in red, the others in blue. In (panel a), we implemented a distance-based approach to point outlier detection. Mahalanobis Distance treats the data as coming from a single distribution, and evaluates single points solely based on their distance from that distribution. As a result, it is unable recognize structured deviations that form coherent groups, even though they are visually distinct.

REDCLAN (panel b) detects outliers based on density, and identifies central portions of outlier sets as inliers, and flags as outliers arbitrary portions of the outer parts of both inlying and outlying regions. This reveals its struggle with varying density patterns, particularly when outlier sets have their own internal density structure different from the main distribution. Clearly, it is not meant to tackle our problem of exclusively finding outlier sets.

CBLOF's effectiveness proves highly dependent on initialization parameters. In panel (c) we show parameters that lead to identify all points in both outlier sets, as well as two smaller sets on the edge of the main inlier set of points. The lack of a verification mechanism for the outlier clusters identified by CBLOF hinders the approach's ability to address our problem of exclusive outlier set identification.

D.MCA (panel d) struggles with coherent outlier set detection due to its point-wise scoring approach. This results in fragmented outlier identification where some points within outlier clusters are detected while others are missed, failing to preserve the structural integrity of outlier sets. Also note that D.MCA shows high sensitivity to hyperparameter choices, and fine-tuning is necessary on a dataset by dataset basis.

Similarly, Feature Bagging (panel e) detects more outliers than our problem warrants, by identifying many outliers all along the dataset's boundaries. This suggests that this method aggregates ensemble model scores in a way that is not compatible for identifying coherent outlier sets.

DPSL (panel f) performs best among existing methods, accurately identifying most points in the

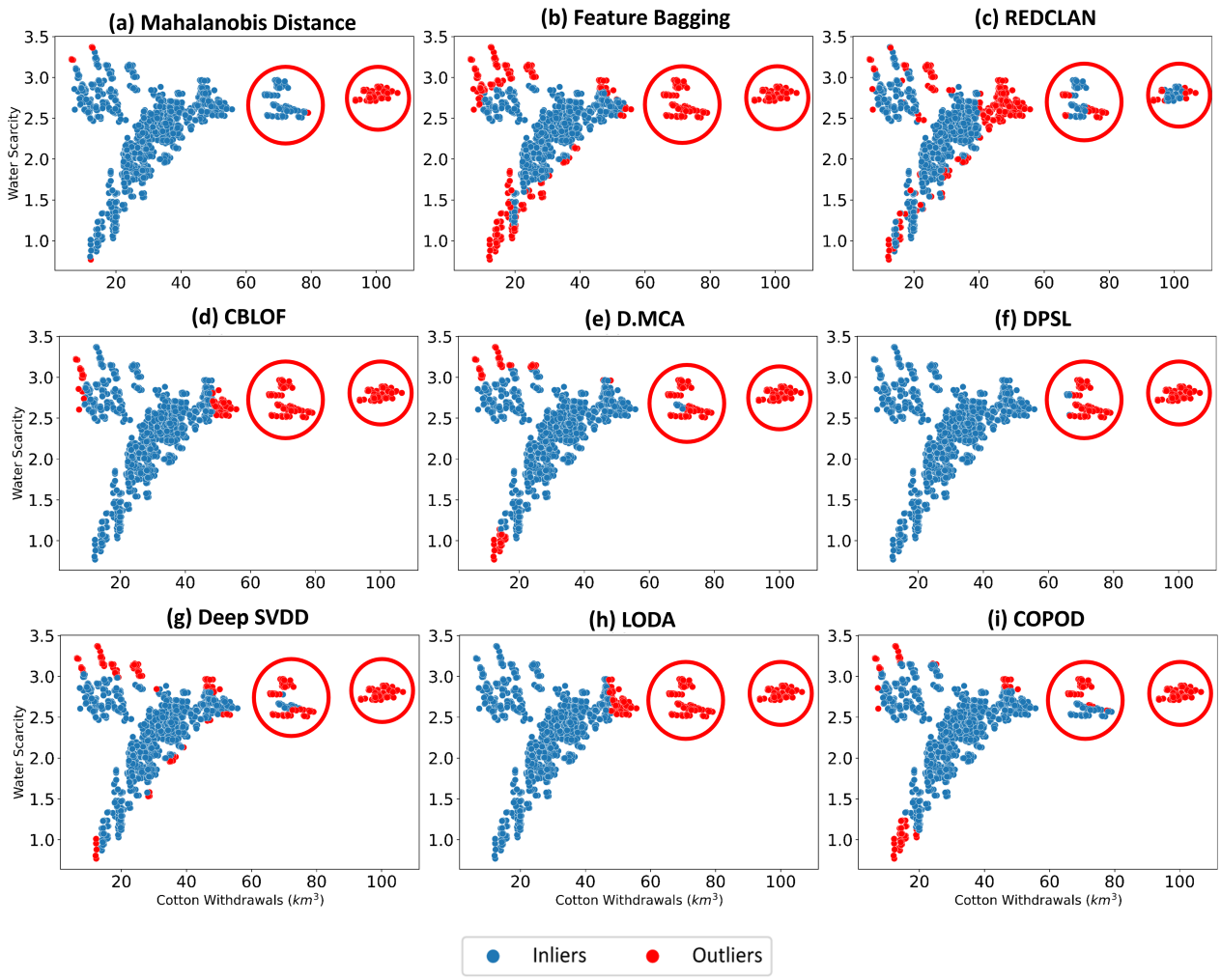

Figure 4: Inliers and outliers as identified by a range of Outlier Detection methods on Figure 1 dataset of cotton irrigation water withdrawals (km^3) vs water scarcity for the Indus River Basin (IRB) in 2100. The benchmark is to identify the two clusters circled in red. (a) Mahalanobis Distance on IRB, (b) Relative-KNN-kernel density-based clustering (REDCLAN) on IRB, (c) Clustering Based Local Outlier Factor (CBLOF) on IRB, (d) Detection with Explicit Micro-Cluster Assignments (D.MCA) on IRB, (e) Feature Bagging on IRB, (f) Differential Potential Spread Loss (DPSL) on IRB, (g)Deep Support Vector Data Description (Deep SVDD) on IRB, (h) Lightweight Online Detector of Anomalies (LODA) on IRB, (i) Copula-Based Outlier Detection (COPOD) on IRB.

outlier sets. However, its effectiveness requires careful parameter tuning and its point-by-point evaluation limits scalability for large datasets where outlier sets need to be identified efficiently.

Deep SVDD (panel g) employs a neural network trained to map input data into a hypersphere of minimum volume in learned feature space, evaluating each point based on its distance to the hypersphere center. Deep SVDD shows scattered outlier detection throughout the dataset, identifying some points within the circled outlier regions but also flagging points elsewhere. This highlights its fundamental limitation: the point-wise evaluation approach prevents recognition of entire clusters as cohesive outlier sets despite identifying some individual points within such regions.

LODA (panel h) aggregates scores from multiple random one-dimensional projections to determine outlier status for each point. LODA exhibits excessive outlier detection, flagging numerous points throughout the dataset including within the circled outlier regions but also multiple false positives. This over-detection occurs because LODA identifies statistical extremes in individual projections rather than recognizing meaningful collective patterns. While it successfully detects points within true outlier clusters, the high false positive rate demonstrates its inability to distinguish between outlier sets and inliers making it unsuitable for exclusive outlier set detection.

COPOD (panel i) analyzes marginal and conditional distributions using copula models to assign outlier scores based on unusual feature dependencies. The method produces scattered outlier identification throughout the IRB dataset, detecting some points within the circled regions but also flagging points elsewhere. This fragmented detection pattern demonstrates how marginal distribution analysis cannot preserve the cluster coherence essential for outlier set identification.

In summary, among all methods tested, DPSL (panel f) performs best on the IRB dataset: it successfully identifies nearly all points in both outlier clusters while maintaining a low false-positive rate among inliers. While it remains sensitive to parameter settings, its relative precision and ability to recover outlier set structure make it the most promising baseline. Accordingly, we carry DPSL forward as a benchmark for subsequent comparative analysis. Our rationale is that if an approach cannot handle the outlier sets (which again, it is not designed to do) in this case where they are obvious, further testing is futile. The results point to the need for an approach specifically designed for outlier set detection with low hyperparameter sensitivity. By designing OSTI as an approach that both preserves cluster structure and identifies outlying behavior, we aim to address these limitations, and next section will now provide evidence of this.

5.2 F1-Score Evaluation

Table 1: Summary of Performance Metrics for Case 1: One outlier set and Case 2: Two outlier sets

Inlier shapes	Ground truth	Case 1: One outlier set					Case 2: Two outlier sets				
		Precision	Recall	F1 score	Purity (%)	Time CPU (sec)	Precision	Recall	F1 score	Purity (%)	Time CPU (sec)
Circle	Strict	0.93	0.96	0.94	99.70	0.27	0.89	0.95	0.92	99.14	0.24
	Relaxed	1.00	0.83	0.91	97.75		0.98	0.87	0.92	97.66	
Ellipse	Strict	0.95	0.87	0.91	99.50	0.23	0.89	0.95	0.92	99.15	0.19
	Relaxed	1.00	0.72	0.83	98.95		0.98	0.85	0.92	98.29	
Triangle	Strict	0.83	1.00	0.91	99.34	0.25	0.87	0.95	0.91	99.18	0.19
	Relaxed	0.97	0.96	0.97	96.15		0.98	0.87	0.92	96.93	
Irregular	Strict	0.89	0.99	0.93	99.81	0.27	0.87	0.95	0.91	99.64	0.24
	Relaxed	1.00	0.87	0.93	97.89		0.98	0.88	0.93	98.21	

Note: The run times in Table 1 are for each 1,000 datasets, with OSTI processing a single dataset in approximately 0.3 seconds on average (or about 3.34 minutes for all 1,000 datasets).

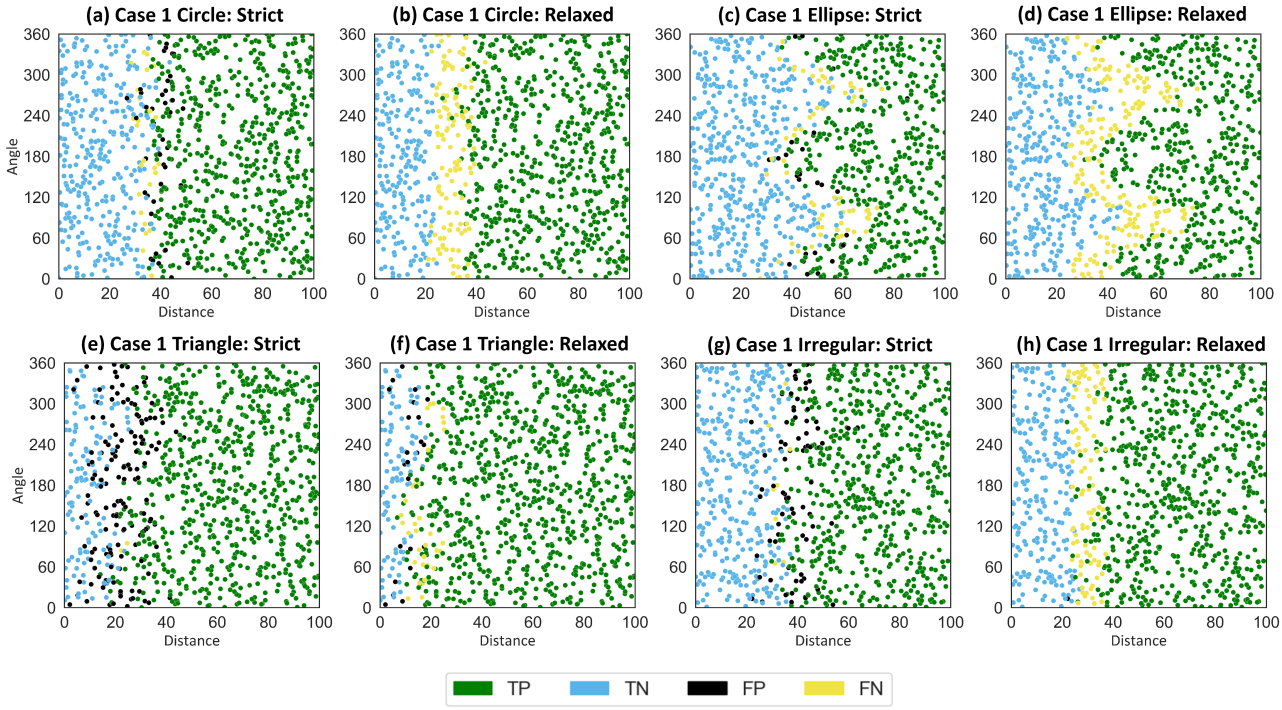

Figure 5: Scatter plot displaying 1,000 datasets across varied distances and angles, illustrating distinct regions for True Positives (TP, in green), True Negatives (TN, in blue), False Positives (FP, in black), and False Negatives (FN, in yellow) for Case 1 with both strict and relaxed ground truth categories. The Figure shows results for: (a-b) Circle, (c-d) Ellipse, (e-f) Triangle, and (g-h) Irregular shapes. Each point represents a single dataset's outcome, plotted based on its distance (x-axis, 0-100 units) and angle (y-axis, 0-360 degrees).

In Table 1, the F1-score is consistently above 0.90 for both Cases 1 and 2. As expected, the strict ground truth decreases the number of false negatives compared with the relaxed ground truth, and increases the number of false positives. This leads to comparatively higher recall and lower precision across all eight experiments in the strict case. In contrast, the relaxed case offers often near perfect precision, highlighting that OSTI detects very few false positives then, i.e., floating sets whose centroid lies within the convex hull of the inlier set. This is an encouraging sign of performance. This also leads to large numbers of false negatives in the relaxed case, culminating in Case 1 with Ellipse inliers.

To visualize this further, Figure 5 illustrates the performance of OSTI on a synthetic dataset shaped as a circle and an ellipse under both strict and relaxed ground truth conditions, in Case 1. For the circle shape, in Figure 5, (a) and (b) the shift from a strict to a relaxed ground truth condition turns most or even all false positives (FP) into true positives (TP), showing the ability of OSTI to detect floating sets whose centroid, but not all points, are outside of the convex hull of the inliers. In the case of the ellipse shape, as shown in Figure 5 (c) and (d), the same trend is observed with the TP increasing under relaxed conditions. However, many floating sets close to the inlier sets are labeled as outlying where OSTI cannot detect them (false negatives, FN, in yellow). This causes the F1 score for ellipse to decreases from 0.91 under strict conditions to 0.83 when relaxed. Note how in both ground truths the separation between positives and negatives varies notably with the angle, showing a zigzag

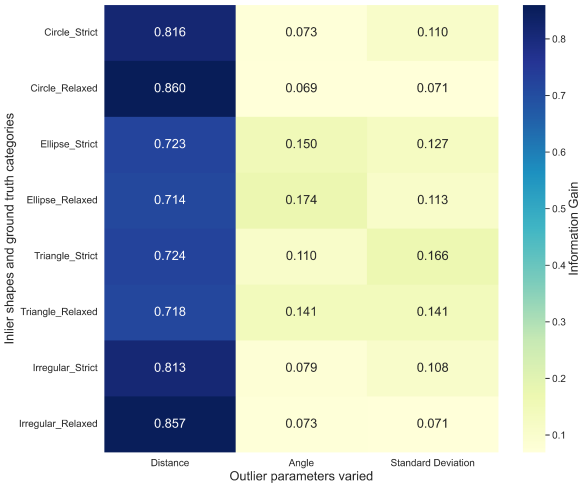

Figure 6: Heatmap showing the relative importance of the three parameters varied across Case 1 (one outlier set) for both strict and relaxed ground truth categories for all four shapes. Information gain was computed using a random forest classifier model, with darker colors representing higher values. Distance is the most influential parameter across all shapes and ground truth categories.

pattern mirroring the shape of the ellipse.

The triangle shape evaluates OSTI's ability to handle sharp geometric features. OSTI achieves F1-scores of 0.91 under strict conditions and improves to 0.97 under relaxed conditions. This notable improvement is evident in Figure 5 (e) and (f), where the strict condition shows a distinctive pattern of FP (black points) clustered particularly around the triangle's vertices. The transition to relaxed ground truth demonstrates how these ambiguous cases are more appropriately classified, as many FP convert to TP (green points) while a smaller number become FN (yellow points) where floating sets partially overlap the inlier set's convex hull. The detection boundaries show clear angular dependencies that correspond to the triangle's corners, illustrating how OSTI naturally adapts its detection criteria to reflect the underlying geometry of the inlier set.

For the irregular shape, OSTI maintains consistent performance with F1-scores of 0.93 across both strict and relaxed conditions, demonstrating its robustness to complex boundary configurations. Figure 5 (g) and (h) represents how the method handles an inlier set with no regular geometric pattern. Under strict conditions, the detection boundary exhibits more fragmentation, directly reflecting the irregular geometry of the inlier set. The transition to relaxed ground truth reveals more continuous detection zones while maintaining perfect precision (1.00), with FN (yellow points) appearing primarily in regions where the complex boundary creates ambiguity.

Thus, results show that ground truth definition can impact OSTI's performance in some cases. Results also highlight how distance (always) and angle (when the inlier is heterotropic) can alone explain detection by OSTI. It is positive to see OSTI's detection ability depend on the inlier shape. To go further, we conducted a parameter importance analysis (see Figure 6) to confirm that 'distance' is more influential followed by 'angle' and 'standard deviation' in detecting outlier sets across all shapes and ground truth categories in Case 1.

We observe similar performance in Case 2, with a F1-score consistently above 0.90. As in Case 1

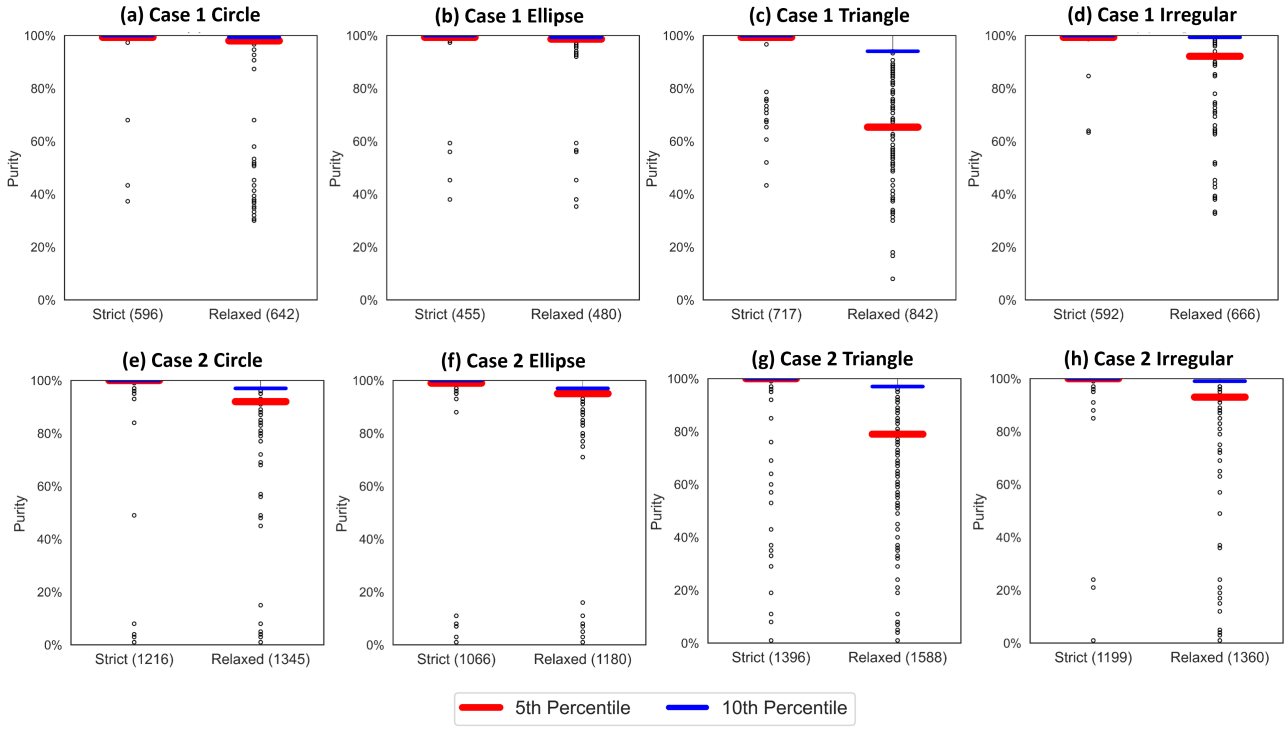

Figure 7: Purity results for True Positives (TP) represented by box plots for Case 1 (one outlier set) (a) Circle, (b) Ellipse, (c) Triangle, (d) Irregular and for Case 2 (Two outlier sets) (e) Circle, (f) Ellipse, (g) Triangle, (h) Irregular representing TP for both ground truth categories. The counts represent the TP identified for each respective ground truth category.

we have higher recall for the strict ground truth, indicating effective identification of TP. Conversely, for the relaxed case, almost perfect precision is achieved, reflecting a minimal rate of FP predictions. These findings indicate that OSTI maintains high accuracy and robustness even in the presence of two outlier sets. More details can be found in Appendix D, Figure 9. Besides, a similar feature analysis to Figure 6 confirmed that 'distance' is the most influential determinant of detection for both outlier sets (Appendix D Figure 10).

5.3 Purity Evaluation

In Table 1, the purity of TP outlier sets is represented and is consistently above 96% for both cases, which implies that OSTI consistently identifies the correct outlier sets. Unsurprisingly, this purity is higher (around or above 99%) under the strict ground truth, where the TP outlier sets are well-separated from the inliers, and the purity decreases under the relaxed ground truth (on average 97%), where some TP outlier sets mix with the inliers.

This distribution of purities across the true positive results is also illustrated by the box plots in Figure 7. We observe that in all cases, around 90% of TP detected by OSTI have perfect purity. Because of this, the full box plot is not visible; instead, the 5th (red) and 10th percentile (blue) lines are represented. For both cases, in Figure7, both percentiles (5th- red line and 10th-blue line), show

that 95% of the datasets have (almost) perfect purity with a strict ground truth, and over 90% with a relaxed ground truth (with the exception of the triangle inlier). Across all four panels, labels indicate the TP counts, with as expected lower counts for strict compared to relaxed ground truth. For Case 1, in Figure 7, panels (a) and (b) show high purity values for over 99% of true positives, illustrating OSTI's ability to correctly identify outlier sets. From Table 1, the strict categories have higher purity (approx. 99.60%) than the relaxed categories (approx. 97.68%) because the strict criteria only count as outlier sets the floating sets that are clearly demarcated from the inliers. This is also evident on Figure 7 panels (a) and (b), which shows more TP values with purity $< 1\%$ identified only under the relaxed ground truth. Panels (c) and (d) reveal different patterns - the triangle shape shows notably lower purity values particularly under relaxed conditions, with more scattered points below 60%, while the irregular shape maintains high purity comparable to circle and ellipse. From Table 1, the strict categories have higher purity (approx. 99.60%) than the relaxed categories (approx. 97.68%) because the strict criteria are more selective. This is evident in all panels, which show more TP values with purity less than 1% identified only under the relaxed ground truth.

For Case 2, in Figure 7, panels (e) and (f), similar to Case 1 shows high purity values, but there is a noticeable increase in outliers as compared to Case 1. Panels (g) and (h) follow similar patterns, the triangle shape again shows more scattered low-purity points, while the irregular shape maintains more consistent high purity values. This increased presence of outliers across all shapes is due to case where the two floating sets overlap in Case 2. Despite this, the high overall purity values demonstrate OSTI's effectiveness in handling multiple outlier sets. Note that the TP counts are nearly double in Case 2 for both ground truths (1216 for strict Circle, 1066 for strict Ellipse, 1396 for strict Triangle, 1199 for strict Irregular and their corresponding relaxed values of 1345, 1180, 1588, and 1360). This doubling of TP counts is a mechanical consequence of the presence of two floating sets per dataset. These findings reinforce the strong performance and reliability of the OSTI method in achieving high purity levels across different inlier and outlier configurations.

5.4 Benchmarking OSTI

Table 2: DPSL results on synthetic dataset for Case 1 & Case 2

Inliers shapes	Case 1 One outlier set		Case 2 Two outlier sets	
	F1-score	Purity (%)	F1-score	Purity (%)
Circle	0.75 ± 0.34	78.77 ± 35.60	0.76 ± 0.23	79.44 ± 35.30
Ellipse	0.60 ± 0.41	63.10 ± 42.88	0.65 ± 0.26	66.90 ± 42.14
Triangle	0.85 ± 0.23	89.46 ± 23.72	0.84 ± 0.17	88.33 ± 25.68
Irregular	0.76 ± 0.33	79.33 ± 34.35	0.75 ± 0.23	79.49 ± 35.08

Next, we implement the best performing method (DPSL) from Section 5.1 on the synthetic datasets used for OSTI's validation. For both Cases 1 and 2 we adopted n: 25, t: 10 and $h_{\max}$: 10, the

same hyperparameter choice as in Section 1, Figure 4 (f). The F1 score and purity for all geometric configurations for both cases are summarized in Table 2. These results show an average F1 score of 0.75 and purity of 79.00% across both cases and all geometric configurations. Each iteration of the analysis took approximately 15 seconds per dataset (258 minutes on average to evaluate 1,000 datasets) and were performed on the same hardware as OSTI. This makes DPSL about 80 times slower than OSTI. The DPSL results are approximately 23% less for F1 score and 26% less for purity as compared with OSTI. This confirms that similar to other approaches aimed at labeling individual outliers, DPSL cannot consistently repeat its performance of Figure 4 (f) without a hyperparameter fine tuning that is inconsistent with the rapid examination of a large number of datasets.

Table 3: Summary of Additional Analysis for Case 1: One outlier set and Case 2: Two outlier sets with Number of clusters=8, Outliers =50 and $\pi_{th} = 0.1$

Inlier shapes	Ground truth	Case 1: One outlier set					Case 2: Two outlier sets				
		Precision	Recall	F1 score	Purity (%)	Time CPU (secs)	Precision	Recall	F1 score	Purity (%)	Time CPU (secs)
Circle	Strict	0.93	1.00	0.96	99.91	0.23	0.95	0.97	0.95	99.86	0.23
	Relaxed	1.00	0.94	0.97	99.20		0.99	0.86	0.92	99.21	
Ellipse	Strict	0.91	0.99	0.95	99.83	0.24	0.95	0.93	0.95	99.79	0.23
	Relaxed	1.00	0.87	0.93	99.38		0.99	0.82	0.89	99.24	
Triangle	Strict	0.86	1.00	0.93	99.98	0.23	0.93	0.95	0.94	99.78	0.23
	Relaxed	0.98	0.99	0.99	98.15		0.98	0.86	0.92	98.39	
Irregular	Strict	0.88	1.00	0.94	99.95	0.18	0.92	0.96	0.94	99.91	0.18
	Relaxed	0.99	0.99	0.99	99.28		0.98	0.88	0.93	99.35	

Table 4: Summary of Additional Analysis for Case 1: One outlier set and Case 2: Two outlier sets with Number of clusters=7, Outliers =150 and $\pi_{th} = 0.1$

Inlier shapes	Ground truth	Case 1: One outlier set					Case 2: Two outlier sets				
		Precision	Recall	F1 score	Purity (%)	Time CPU (secs)	Precision	Recall	F1 score	Purity (%)	Time CPU (secs)
Circle	Strict	0.94	0.96	0.95	99.78	0.23	0.90	0.95	0.93	99.69	0.21
	Relaxed	1.00	0.82	0.90	99.42		0.99	0.86	0.92	97.25	
Ellipse	Strict	0.95	0.87	0.91	99.84	0.20	0.90	0.8	0.92	99.49	0.22
	Relaxed	1.00	0.72	0.84	99.60		0.98	0.85	0.91	98.96	
Triangle	Strict	0.84	1.00	0.91	99.41	0.18	0.88	0.95	0.92	99.42	0.20
	Relaxed	0.98	0.95	0.97	96.99		0.98	0.86	0.92	95.68	
Irregular	Strict	0.90	0.99	0.94	99.96	0.13	0.88	0.95	0.92	99.72	0.24
	Relaxed	1.00	0.86	0.93	98.95		0.98	0.88	0.93	96.98	

Table 5: Summary of Additional Analysis for Case 1: One outlier set and Case 2: Two outlier sets with Number of clusters=9, Outliers =150 and $\pi_{th} = 0.1$

Inlier shapes	Ground truth	Case 1: One outlier set					Case 2: Two outlier sets				
		Precision	Recall	F1 score	Purity (%)	Time CPU (secs)	Precision	Recall	F1 score	Purity (%)	Time CPU (secs)
Circle	Strict	0.92	0.96	0.94	99.69	0.21	0.90	0.95	0.92	99.14	0.21
	Relaxed	1.00	0.84	0.91	97.25		0.99	0.87	0.92	97.59	
Ellipse	Strict	0.95	0.87	0.91	99.49	0.20	0.90	0.94	0.92	99.15	0.24
	Relaxed	1.00	0.71	0.83	98.96		0.99	0.85	0.91	98.27	
Triangle	Strict	0.82	1.00	0.90	99.42	0.21	0.88	0.95	0.92	99.08	0.21
	Relaxed	0.97	0.96	0.97	95.68		0.98	0.87	0.93	96.52	
Irregular	Strict	0.87	0.99	0.93	99.72	0.22	0.88	0.96	0.92	99.62	0.21
	Relaxed	1.00	0.89	0.94	96.98		0.98	0.88	0.93	98.04	

5.5 Hyperparameter Robustness Checks

In contrast to DPSL, with a single hyperparameter choice, OSTI achieves a robustly high and consistent level of accuracy in detecting outlier sets across various geometric configurations, with a computational efficiency that is two orders of magnitude better. It maintains robust outlier detection whether the outlier set size is similar to the weight threshold (Case 1) or half the weight threshold (Case 2). To provide further evidence that outlier set detection has low sensitivity to its weight being well-below the weight threshold, we also checked that keeping the same weight threshold with outlier set weight divided by three does not affect OSTI performance (Table 3), maintaining high F1 scores (average 0.94) and purity (average 99.38%) with consistent computational efficiency (0.2 secs per dataset). This suggests that setting the weight threshold at the maximum weight at which one will consider a set to be small enough an outlier set (and not part of the inliers) is a choice that enables to capture outlier sets of all sizes. Our experiments therefore indicate that a weight threshold such as $\pi_{th} = 0.1$ (10% outliers), relating to a value of K slightly smaller than $1/\pi_{th} = 10$, is a sensible choice for OSTI hyperparameters. We also verified that OSTI is robust to the exact choice of K, with $K = 7$ (Table 4: average F1 score 0.92, average purity 98.86%, 3.31 minutes per 1000 datasets) and $K = 9$ (Table 5: average F1 score 0.91, average purity 98.11%, 3.46 minutes per 1000 datasets) delivering results close to those of Table 1. In practice, setting the optimal threshold will depends on the specific characteristics of a dataset and on the desired sensitivity to outliers.

The weight threshold π_{th} is a key hyperparameter as it determines the maximum proportion of data points that can constitute an outlier set. Recall that over this threshold, a cluster is considered part of the main data distribution and not a potentially outlying cluster. In practice, π_{th} should be set based on domain knowledge and the specific characteristics of the dataset being analyzed. For exploratory data analysis where the goal is to identify meaningful deviations from typical patterns, we recommend $\pi_{\text{th}} = 0.1$ (10% of the dataset) as a robust starting point. This value ensures that clusters representing more than 10% of the data are considered part of the main distribution rather than outliers, which aligns with conventional outlier detection principles where outliers represent a small fraction of the data. The choice of π_{th} is intrinsically linked to the number of clusters K in the GMM. To ensure effective outlier set identification, we recommend choosing K such that $1/K$ is slightly larger than π_{th}, allowing the method to detect clusters smaller than the average cluster size. For instance, with $\pi_{\text{th}} = 0.1$, setting $K = 8$ yields an average cluster weight of 0.125, creating the necessary sensitivity to identify outlier sets. Our extensive robustness analysis demonstrates that OSTI maintains consistent performance across different threshold values and cluster numbers ($K = 7, 8, 9$), with F1 scores remaining above 0.90 and purity exceeding 98% even when outlier sets are significantly smaller than the threshold. This robustness eliminates the need for precise parameter tuning, making OSTI practical for rapid analysis of large datasets. For datasets where smaller anomalous groups are expected, π_{th} can be reduced (e.g., 0.05 for 5% threshold), while larger thresholds (e.g., 0.15) may be appropriate when focusing on more substantial deviations from normal patterns.

6 Discussion

OSTI fills a gap by providing a means to detect an outlier cluster in a large dataset with a single label, while excluding individual outlying points from detection. State-of-the-art outlier detection methods tested in this paper are not designed to tackle this problem. Indeed, they are unable to detect outlier sets without also flagging individual outlying points. In contrast, OSTI's two-step approach combining GMM clustering with statistical verification provides an effective solution for detecting cohesive outlier sets.

Our approach addresses a real-world need as demonstrated through the IRB climate scenario analysis, where coherent anomalous patterns cannot be identified in isolation through traditional outlier detection methods. This makes the insights from outlier set detection unattainable through individual point analysis. However, existing benchmark repositories are designed for individual point detection and lack established ground truth for collective outlier patterns, creating a methodological gap for evaluating outlier set detection approaches. Therefore, our synthetic dataset approach provides methodological advantages that would be impossible to achieve with existing real-world datasets. First, it enables rigorous evaluation with known ground truth, allowing us to measure precise metrics such as purity, which quantifies how accurately our detected outlier sets correspond to the true synthetically generated outlier sets. Second, it allows systematic exploration of diverse scenarios across 8,000 datasets with varying inlier geometries and outlier set characteristics, providing comprehensive evidence of OSTI's robustness. Third, it enables controlled experimentation where we can isolate specific factors influencing performance, such as distance, angle, and cluster spread. Our validation across these synthetic datasets demonstrates OSTI's robust performance with F1 scores consistently above 0.90 and average purity of 98.58% for detected outlier sets.

A next step is the development of specialized benchmark datasets for outlier set detection. It is beyond the scope of this work because there is not yet a need to compare methods for the exclusive detection of outlier sets. This follows an established precedent in anomaly detection, where foundational methods are typically introduced well before systematic benchmarks are constructed to compare a growing number of approaches. For example, classic techniques such as Grubbs' procedures (1969) (*Grubbs*, 1969) and Mahalanobis distance (1930s) (*Mahalanobis*, 1933) were developed decades prior to the release of the first comprehensive benchmark repository in 2016 (*Campos et al.*, 2016c). Our synthetic dataset approach aligns with this pattern, providing the necessary methodological foundation, evaluation criteria, and characterization of outlier sets that can inform the eventual construction of real-world benchmarks for this newly formalized problem.

Even though results showcased the superior computational efficiency of OSTI compared with DPSL, these points should be discussed in more detail. The GMM step operates on the entire dataset of n points with a computational complexity of order $O(I \cdot n \cdot K \cdot d^2)$ (*Chivers and Sleightholme*, 2015), where I is the number of EM iterations, K the number of clusters, and d the data dimensionality. In contrast, the Mahalanobis distance computation is applied only to the K cluster centroids and thus requires significantly fewer operations, of order $O(K \cdot d^2)$. Therefore, OSTI's computational time is dominated by the GMM clustering step rather than the inter-cluster Mahalanobis distance computation. Given that K is typically small between 7 and 9 compared to n, which ranges from hundreds to thousands, the GMM step dominates the overall runtime. In practical terms, with parameters $n = 1500$, $d = 2$, $K = 8$,

and $I = 20$, the number of operations remains under a million and yields an empirical runtime of 0.3 seconds per dataset on standard hardware. This computational profile scales linearly with dataset size, cluster number, and EM iterations, and quadratically with dimensionality due to covariance operations. Compared to DPSL, which exhibits quadratic complexity in n, OSTI's linear scaling and low runtime enable it to process datasets approximately 50 times faster without compromising detection accuracy, making it well-suited for exploratory tasks involving multiple low-dimensional projections.

The evaluation we provide for OSTI focuses on two-dimensional analysis. A limitation of our approach is that the chi-square test used in OSTI's second step loses statistical power as the number of degrees of freedom increases (*Fisher*, 1922; *Wilson and Hilferty*, 1931), and this could compromise OSTI's ability to reliably detect outlier sets in higher dimensions. However, this apparent limitation can be overcome by breaking complex high-dimensional problems into multiple interpretable low-dimensional analyses. Rather than attempting to detect outlier sets in high-dimensional space where relationships become increasingly difficult to interpret. OSTI enables systematic exploration of low-dimensional projections where patterns remain interpretable and statistically verifiable.

If OSTI were to be scaled to higher dimensions, several refinements should be considered, including: (1) Dimension reduction techniques like Principal Component Analysis as a preprocessing step (*Reddy et al.*, 2020; *Fodor*, 2002), (2) Enhanced GMM initialization and convergence criteria to handle high-dimensional spaces efficiently (*Zhao et al.*, 2018), and (3) Alternatives to the chi-square test that maintain statistical power in higher dimensions e.g., (*Hotelling et al.*, 1931; *Holloway and Dunn*, 1967; *Rousseeuw and Driessen*, 1999). However, these modifications would involve trade-offs with OSTI's current advantages in low-dimensional settings, where it provides clear benefits: statistical tests maintain their power and reliability, computational efficiency is preserved (0.3 seconds per dataset), results can be easily visualized and interpreted, and the method can be systematically applied across multiple low-dimensional projections of complex data.

The hyperparameter sensitivity of OSTI, particularly regarding the weight threshold (π_{th}) and number of clusters (K), warrants careful examination. The weight threshold π_{th} fundamentally determines what proportion of points can constitute an outlier set – setting this parameter too high risks missing genuinely outlying patterns, while setting it too low may lead to excessive fragmentation of clusters. Our empirical testing suggests $\pi_{th} = 0.1$ provides a robust balance, successfully detecting outlier sets while maintaining high purity (98.58%) across diverse scenarios.

The number of clusters K interacts with π_{th} in important ways. Theoretically, K must be large enough to allow identification of small outlier clusters, yet small enough to avoid over-segmentation of the data. We found that setting K slightly below $1/\pi_{th}$ (e.g., $K = 8$ for $\pi_{th} = 0.1$) consistently produces reliable results. This relationship is intuitive: if K is too small relative to $1/\pi_{th}$, then the average cluster size would exceed the outlier threshold, potentially preventing detection of genuine outlier sets.

Remarkably, our evaluation demonstrates that OSTI results remain consistent across different clustering parameters ($K = 7, 8, 9$), and even when outlier sets are significantly smaller (one-third the size) than the weight threshold. This robustness to parameter choices indicates that OSTI can reliably detect outlier sets of varying sizes without requiring precise parameter tuning - a significant practical advantage compared to methods like DPSL that require careful hyperparameter optimization. Future

work could explore multi-scale detection frameworks that identify outlier patterns at different granularities, streaming applications for real-time outlier set monitoring, and domain-specific validations in fields requiring collective anomaly understanding (*Jiang et al.*, 2019).

7 Conclusion

This paper introduces a new type of outlier, the *outlier set* which is defined as a cluster of points that deviates significantly from the rest of the dataset, and that we consider and evaluate as a single entity. This new term addresses an emerging gap in outlier detection – as datasets grow larger and more complex, identifying and interpreting individual outlying points becomes impractical compared to detecting meaningful outlying patterns involving groups of points.

Existing outlier detection methods focus on detecting individual outlier points, and are therefore not trained to consistently detect these outlier sets while excluding isolated outlying points. To address this gap, we propose a new methodology, Outlier Set Two-step Identification (OSTI), which combines Gaussian Mixture Models (GMM), Inter-cluster Mahalanobis distance (IMD), and chi-square based hypothesis testing. We evaluate OSTI on 8,000 synthetic datasets featuring varied inlier geometries and outlier set characteristics. This extensive validation demonstrates OSTI's effectiveness in 2D spaces, where it consistently achieves F1 scores above 0.90 and maintains high purity (98.58%) across diverse configurations. A key area for future work is to evaluate OSTI's performance higher dimensions, where the theoretical power of its statistical testing is set to decrease, and where interpretability and efficiency might become challenges. This would also require careful consideration of how the GMM-based clustering and statistical verification steps scale with dimension. Alternative distance metrics and dimension reduction techniques could help address the curse of dimensionality while preserving OSTI's ability to identify meaningful pattern deviations.

OSTI has demonstrable potential for applications in multiple domains, such as anomaly detection in large-scale networks, healthcare, bio-informatics, and climate risk modeling, to name a few. In all these fields, the ability to efficiently detect and analyze groups of outlying points as cohesive sets rather than individual outliers could provide crucial insights for decision-making.

Acknowledgements

We are grateful to Commonwealth Scholarship Commission, Grantham Center for Sustainable Futures and University of Sheffield Postgraduate Researcher Publication Scholarships for funding Dr Amal Sarfraz's PhD research. Dr Charles Rougé is partially supported by the UK Engineering and Physical Sciences Research Council through the "Flexible design and operation of water resource systems to tackle the triple challenge of climate change, the energy transition, and population growth" project (Ref: EP/X009459/1). We also want to thank the four anonymous reviewers and the associate editor whose comments helped markedly improve this work.

CRediT authorship contribution statement

Amal Sarfraz: Conceptualization, Methodology, Formal analysis, Software, Investigation, Visualization, Validation, Data curation, Writing - original draft, Writing - review & editing. **Abigail Birnbaum**: Software, Resources, Writing - review & editing. **Flannery Dolan:** Resources, Writing - review & editing. **Jonathan Lamontagne**: Resources, Writing - review & editing. **Lyudmila Mihaylova**: Conceptualization, Visualization, Supervision, Writing - review & editing. **Charles Rougé**: Conceptualization, Visualization, Validation, Resources, Supervision, Writing - review & editing.

References

Aggarwal, C. C., and S. Sathe (2015), Theoretical foundations and algorithms for outlier ensembles, *ACM SIGKDD Explorations Newsletter*, *17*, 24–47, doi:10.1145/2830544.2830549.

Allenby, M. C., E. S. Liang, J. Harvey, M. A. Woodruff, M. Prior, C. D. Winter, and D. Alonso-Caneiro (2021), Detection of clustered anomalies in single-voxel morphometry as a rapid automated method for identifying intracranial aneurysms, *Computerized Medical Imaging and Graphics*, *89*, 101,888, doi:10.1016/j.compmedimag.2021.101888.

Bai, M., X. Wang, J. Xin, and G. Wang (2016), An efficient algorithm for distributed density-based outlier detection on big data, *Neurocomputing*, *181*, 19–28, doi:10.1016/j.neucom.2015.05.135.

Barnett, V. (1978), The study of outliers: purpose and model, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *27*(3), 242–250.

Barz, B., Y. G. Garcia, E. Rodner, and J. Denzler (2017), Maximally divergent intervals for extreme weather event detection, in *OCEANS 2017-Aberdeen*, pp. 1–9, IEEE, doi:10.1109/OCEANSE.2017.8084569.

Basak, S., A. Ayman, A. Laszka, A. Dubey, and B. Leao (2019), Data-driven detection of anomalies and cascading failures in traffic networks, in *Annual Conference of the PHM Society 2019*, vol. 11, doi:10.36001/phmconf.2019.v11i1.861.

Beckman, R., and R. Cook (1983), Outlier â¦ â¦ â¦. s, *Technometrics*, *25*(2), 119–149, doi:10.1080/00401706.1983.10487840.

Benabderrahmane, S., N. Hoang, P. Valtchev, J. Cheney, and T. Rahwan (2024), Hack me if you can: Aggregating autoencoders for countering persistent access threats within highly imbalanced data, *Future Generation Computer Systems*, *160*, 926–941, doi:10.48550/arXiv.2406.19220.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, and J. Sander (2000), Lof: Identifying density-based local outliers, in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, doi:10.1145/342009.335388.

Byers, E., V. Krey, E. Kriegler, K. Riahi, R. Schaeffer, J. Kikstra, R. Lamboll, Z. Nicholls, M. Sandstad, C. Smith, et al. (2022), Ar6 scenarios database, *Tech. rep.*, International Institute for Applied Systems Analysis.

Calvin, K., B. Bond-Lamberty, L. Clarke, J. Edmonds, J. Eom, C. Hartin, S. Kim, P. Kyle, R. Link, R. Moss, et al. (2017), The ssp4: A world of deepening inequality, *Global Environmental Change*, *42*, 284–296, doi:10.1016/j.gloenvcha.2016.06.010.

Calvin, K., P. Patel, L. Clarke, G. Asrar, B. Bond-Lamberty, R. Y. Cui, A. Di Vittorio, K. Dorheim, J. Edmonds, C. Hartin, et al. (2019), Gcam v5. 1: representing the linkages between energy, water, land, climate, and economic systems, *Geoscientific Model Development*, *12*(2), 677–698, doi:10.5194/gmd-12-677-2019.

Campos, G. O., A. Zimek, J. Sander, R. J. G. B. Campello, B. MicenkovÃ¡, E. Schubert, I. Assent, and M. E. Houle (2016a), On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study, *Data Mining and Knowledge Discovery*, *30*, 891–927, doi:10.1007/s10618-015-0444-8.

Campos, G. O., A. Zimek, J. Sander, R. J. G. B. Campello, B. MicenkovÃ¡, E. Schubert, I. Assent, and M. E. Houle (2016b), Supplementary material for on the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study, [`https://elki-project.github.io/datasets/outlier`], accessed: 2023-07-23.

Campos, G. O., A. Zimek, J. Sander, R. J. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle (2016c), On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study, *Data mining and knowledge discovery*, *30*, 891–927, doi:10.1007/s10618-015-0444-8.

Cheong, M.-Y., and H. Lee (2008), Determining the number of clusters in cluster analysis, *Journal of the Korean Statistical Society*, *37*, 135–143.

Chinchor, N., and B. Sundheim (1993), Muc-5 evaluation metrics, in *Proceedings of the 5th Conference on Message Understanding*, MUC5 '93, p. 69â78, Association for Computational Linguistics, USA, doi:10.3115/1072017.1072026.

Chivers, I., and J. Sleightholme (2015), *An Introduction to Algorithms and the Big O Notation*, pp. 359–364, Springer International Publishing, Cham, doi:10.1007/978-3-319-17701-4_23.

Coelho, C., C. Ferro, D. Stephenson, and D. Steinskog (2008), Methods for exploring spatial and temporal variability of extreme events in climate data, *Journal of Climate*, *21*(10), 2072–2092, doi:10.1175/2007JCLI1781.1.

Davies, D. L., and D. W. Bouldin (1979), A cluster separation measure, *IEEE transactions on pattern analysis and machine intelligence*, *2*, 224–227, doi:10.1109/TPAMI.1979.4766909.

Day, N. E. (1969), Estimating the components of a mixture of normal distributions, *Biometrika*, *56*(3), 463–474.

Dekker, M. M., A. F. Hof, M. van den Berg, V. Daioglou, R. van Heerden, K.-I. van der Wijst, and D. P. van Vuuren (2023), Spread in climate policy scenarios unravelled, *Nature*, *624*(7991), 309–316, doi:10.1038/s41586-023-06738-6.

Divya, D., and S. S. Babu (2016), Methods to detect different types of outliers, in *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pp. 23–28, IEEE, doi:10.1109/SAPIENCE.2016.7684114.

Dolan, F., J. Lamontagne, R. Link, M. Hejazi, P. Reed, and J. Edmonds (2021), Evaluating the economic impact of water scarcity in a changing world, *Nature Communications*, *12*, 1915, doi:10.1038/s41467-021-22194-0.

Emmott, A., S. Das, T. Dietterich, A. Fern, and W.-K. Wong (2015), A meta-analysis of the anomaly detection problem, *arXiv: Artificial Intelligence*.

Emmott, A. F., S. Das, T. Dietterich, A. Fern, and W. K. Wong (2013), Systematic construction of anomaly detection benchmarks from real data, in *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description, ODD 2013*, pp. 16–21, doi:10.1145/2500853.2500858.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (1996), A density-based algorithm for discovering clusters in large spatial databases with noise, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, p. 226â231, AAAI Press.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor (2016), Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization, *Geoscientific Model Development*, *9*(5), 1937–1958, doi:10.5194/gmd-9-1937-2016.

Feroze, A., A. Daud, T. Amjad, and M. K. Hayat (2021), Group anomaly detection: past notions, present insights, and future prospects, *SN Computer Science*, *2*, 1–27, doi:10.1007/s42979-021-00603-x.

Fisher, R. A. (1922), On the interpretation of χ 2 from contingency tables, and the calculation of p, *Journal of the royal statistical society*, *85*(1), 87–94, doi:10.2307/2340521.

Fodor, I. K. (2002), A survey of dimension reduction techniques, *Tech. rep.*, Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States).

Gao, X., J. Yu, S. Zha, S. Fu, B. Xue, P. Ye, Z. Huang, and G. Zhang (2022), An ensemble-based outlier detection method for clustered and local outliers with differential potential spread loss, *Knowledge-Based Systems*, *258*, doi:10.1016/j.knosys.2022.110003.

Ghorbani, H. (2019), Mahalanobis distance and its applications for detecting multivariate outliers, *Facta Universitatis, Series: Mathematics and Informatics*, p. 583, doi:10.22190/fumi1903583g.

Grubbs, F. E. (1969), Procedures for detecting outlying observations in samples, *Technometrics*, *11*(1), 1–21.

Grubbs, F. E., and G. Beck (1972), Extension of sample sizes and percentage points for significance tests of outlying observations, *Technometrics*, *14*(4), 847–854, doi:10.1080/00401706.1972.10488981.

Gupta, R., B. Tian, Y. Wang, and K. Nahrstedt (2024), Twin-adapt: Continuous learning for digital twin-enabled online anomaly classification in iot-driven smart labs, *Future Internet*, *16*(7), 239, doi:10.3390/fi16070239.

Hawkins, D. M. (1980), *Multivariate outlier detection*, 104–114 pp., Springer, doi:10.1007/978-94-015-3994-4_8.

He, Z., X. Xu, and S. Deng (2003), Discovering cluster-based local outliers, *Pattern Recognition Letters*, *24*, 1641–1650, doi:10.1016/S0167-8655(03)00003-5.

Hodge, V. J., and J. Austin (2004), A survey of outlier detection methodologies, *Artificial Intelligence Review*, *22*, 85–126, doi:10.1023/B:AIRE.0000045502.10941.a9.

Holloway, L. N., and O. J. Dunn (1967), The robustness of hotelling's t 2, *Journal of the American Statistical Association*, *62*(317), 124–136, doi:10.2307/2282915.

Hotelling, H., et al. (1931), The generalization of student's ratio, *The Annals of Mathematical Statistics*.

Ishioka, T., et al. (2005), An expansion of x-means for automatically determining the optimal number of clusters, in *Proceedings of International Conference on Computational Intelligence*, vol. 2, pp. 91–95.

Jiang, F., G. Liu, J. Du, and Y. Sui (2016), Initialization of k-modes clustering using outlier detection techniques, *Information Sciences*, *332*, 167–183, doi:10.1016/j.ins.2015.11.005.

Jiang, F., H. Zhao, J. Du, Y. Xue, and Y. Peng (2019), Outlier detection based on approximation accuracy entropy, *International Journal of Machine Learning and Cybernetics*, *10*, 2483–2499, doi:10.1007/s13042-018-0884-8.

Jiang, S., R. L. F. Cordeiro, and L. Akoglu (2022), D.mca: Outlier detection with explicit micro-cluster assignments, *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 987–992.

Kasture, P., and J. Gadge (2012), Cluster based outlier detection, *International Journal of Computer Applications*, *58*, 11–15, doi:10.5120/9317-3549.

Kelly, M., R. Longjohn, and K. Nottingham (2017), The UCI machine learning repository.

Kriegel, H.-P., M. Schubert, and A. Zimek (2008), Angle-based outlier detection in high-dimensional data, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, p. 444â452, Association for Computing Machinery, New York, NY, USA, doi:10.1145/1401890.1401946.

Kwakkel, J. H., and E. Pruyt (2013), Exploratory modeling and analysis, an approach for model-based foresight under deep uncertainty, *Technological Forecasting and Social Change*, *80*(3), 419–431, doi:10.1016/j.techfore.2012.10.005.

Lamontagne, J. R., P. M. Reed, R. Link, K. V. Calvin, L. E. Clarke, and J. A. Edmonds (2018), Large ensemble analytic framework for consequence-driven discovery of climate change scenarios, *Earth's Future*, *6*(3), 488–504, doi:10.1002/2017EF000701.

Lazarevic, A., and V. Kumar (2005), Feature bagging for outlier detection, in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 157–166, doi:10.1145/1081870.1081891.

Lee, K., Y. Jeong, S. Joo, Y. S. Yoon, S. Han, and H. Baik (2022), Outliers in financial time series data: outliers, margin debt, and economic recession, *Machine Learning with Applications*, *10*, 100,420, doi:10.1016/j.mlwa.2022.100420.

Li, K., X. Gao, X. Jia, B. Xue, S. Fu, Z. Liu, X. Huang, and Z. Huang (2022), Detection of local and clustered outliers based on the density–distance decision graph, *Engineering Applications of Artificial Intelligence*, *110*, 104,719, doi:10.1016/j.engappai.2022.104719.

Li, X., S. Deng, L. Li, and Y. Jiang (2019), Outlier detection based on robust mahalanobis distance and its application, *Open Journal of Statistics*, *9*(1), 15–26, doi:10.4236/ojs.2019.91002.

Li, Z., and L. Zhang (2023), An ensemble outlier detection method based on information entropy-weighted subspaces for high-dimensional data, *Entropy*, *25*(8), 1185, doi:10.3390/e25081185.

Li, Z., Y. Zhao, N. Botta, C. Ionescu, and X. Hu (2020), Copod: copula-based outlier detection, in *2020 IEEE international conference on data mining (ICDM)*, pp. 1118–1123, IEEE, doi:10.48550/arXiv.2009.09463.

Liu, F. T., K. M. Ting, and Z.-H. Zhou (2008), Isolation forest, in *2008 eighth ieee international conference on data mining*, pp. 413–422, IEEE, doi:10.1109/ICDM.2008.17.

Madabhushi, S., and R. Dewri (2023), A survey of anomaly detection methods for power grids, *International Journal of Information Security*, *22*(6), 1799–1832, doi:10.1007/s10207-023-00720-z.

Maesschalck, R. D., D. Jouan-Rimbaud, and D. L. Massart (2000), The mahalanobis distance, *Chemometrics and Intelligent Laboratory Systems*, *50*, 1–18, doi:10.1016/S0169-7439(99)00047-7.

Mahalanobis, P. C. (1933), On tests and measures of group divergence, *Journal and Proceedings of the Asiatic Society of Bengal*, *26*, 541–588.

Mandhare, H. C., and S. R. Idate (2017), A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques, in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 931–935, IEEE, doi:10.1109/ICCONS.2017.8250601.

Mazzarisi, P., A. Ravagnani, P. Deriu, F. Lillo, F. Medda, and A. Russo (2024), A machine learning approach to support decision in insider trading detection, *EPJ Data Science*, *13*(1), 66, doi:10.2139/ssrn.4294752.

McKay, M. D., R. J. Beckman, and W. J. Conover (2000), A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics*, *42*(1), 55–61, doi:10.2307/1268522.

Moallemi, E. A., J. Kwakkel, F. J. de Haan, and B. A. Bryan (2020), Exploratory modeling for analyzing coupled human-natural systems under uncertainty, *Global Environmental Change*, *65*, 102,186, doi:10.1016/j.gloenvcha.2020.102186.

Ou, Y., C. Roney, J. Alsalam, K. Calvin, J. Creason, J. Edmonds, A. A. Fawcett, P. Kyle, K. Narayan, P. OâRourke, et al. (2021), Deep mitigation of CO2 and non-CO2 greenhouse gases toward 1.5Â° c and 2Â° c futures, *Nature Communications*, *12*(1), 6245, doi:10.1038/s41467-021-26509-z.

Ouyang, B., Y. Song, Y. Li, G. Sant, and M. Bauchy (2021), Ebod: An ensemble-based outlier detection algorithm for noisy datasets, *Knowledge-Based Systems*, *231*, 107,400, doi:10.1016/j.knosys.2021.107400.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011), Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, *12*, 2825–2830.

Pevnỳ, T. (2016), Loda: Lightweight on-line detector of anomalies, *Machine Learning*, *102*, 275–304, doi:10.1007/s10994-015-5521-0.

Powers, D. (2011), Evaluation: From precision, recall and f-measure to roc, informedness, markedness correlation, *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Ramachandra, B., K. K. Gadiraju, R. R. Vatsavai, D. P. Kaiser, and T. P. Karnowski (2016), Detecting extreme events in gridded climate data, *Procedia Computer Science*, *80*, 2397–2401, doi:10.1016/j.procs.2016.05.537.

Ramaswamy, S., R. Rastogi, and K. Shim (2000), Efficient algorithms for mining outliers from large data sets, *ACM SIGMOD Record*, *29*, 427–438, doi:10.1145/335191.335437.

Rayana, S. (2016), ODDS library.

Reddy, G. T., M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker (2020), Analysis of dimensionality reduction techniques on big data, *Ieee Access*, *8*, 54,776–54,788, doi:10.1109/ACCESS.2020.2980942.

Rousseeuw, P. J. (1987), Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics*, *20*, 53–65, doi:10.1016/0377-0427(87)90125-7.

Rousseeuw, P. J., and K. V. Driessen (1999), A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, *41*(3), 212–223, doi:10.2307/1270566.

Ruff, L., R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft (2018), Deep one-class classification, in *Proceedings of the 35th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, vol. 80, pp. 4393–4402, PMLR.

Saha, D., D. Banerjee, and B. P. Majumder (2018), Redclan-relative density based clustering and anomaly detection, *Computer Science & Information Technology (CS & IT)*, pp. 25–39, doi:10.5121/csit.2018.81303.

Sahu, S. K., D. P. Mohapatra, and N. K. Ray (2021), An ensemble-based outlier detection approach on intrusion detection, in *Proceedings of the 19th OITS International Conference on Information Technology (OCIT)*, pp. 404–409, IEEE, doi:10.1109/OCIT53463.2021.00085.

Sánchez Vinces, B. V., E. Schubert, A. Zimek, and R. L. Cordeiro (2025), A comparative evaluation of clustering-based outlier detection, *Data Mining and Knowledge Discovery*, *39*(2), 13, doi:10.1007/s10618-024-01086-z.

Sarfraz, A., A. Birnbaum, F. Dolan, J. Lamontagne, L. Mihaylova, and C. Rouge (2025), Outlier set two-step method (OSTI), Dataset, doi:10.15131/shef.data.28227974.v2.

Sasaki, Y., et al. (2007), The truth of the f-measure, *Teach tutor mater*, *1*(5), 1–5.

Schwarz, G. (1978), Estimating the dimension of a model, *The Annals of Statistics*, *6*(2), 461–464.

Singh, K., and S. Upadhyaya (2012), Outlier detection: Applications and techniques, *International Journal of Computer Science Issues*, *9*, 307–323.

Steinbuss, G., and K. Böhm (2021), Benchmarking unsupervised outlier detection with realistic synthetic data, *ACM Transactions on Knowledge Discovery from Data*, *15*, doi:10.1145/3441453.

Thudumu, S., P. Branch, J. Jin, and J. Singh (2020), A comprehensive survey of anomaly detection techniques for high dimensional big data, *Journal of big data*, *7*, 1–30, doi:10.1186/s40537-020-00320-x.

Van Rijsbergen, C. J. (1977), A theoretical basis for the use of co-occurrence data in information retrieval, *Journal of documentation*, *33*(2), 106–119.

Wilson, E. B., and M. M. Hilferty (1931), The distribution of chi-square, *Proceedings of the National Academy of Sciences*, *17*(12), 684–688.

Yu, R., X. He, and Y. Liu (2015), Glad: Group anomaly detection in social media analysis, *ACM Trans. Knowl. Discov. Data*, *10*(2), doi:10.1145/2811268.

Zhao, Y., A. K. Shrivastava, and K. L. Tsui (2018), Regularized Gaussian mixture model for high-dimensional clustering, *IEEE transactions on cybernetics*, *49*(10), 3677–3688, doi:10.1109/TCYB.2018.2846404.

Zhao, Y., Z. Nasrullah, and Z. Li (2019), Pyod: A python toolbox for scalable outlier detection.

Zimek, A., R. J. Campello, and J. Sander (2014), Ensembles for unsupervised outlier detection: challenges and research questions a position paper, *SIGKDD Explorations*, *15*(1), doi:10.1145/2594473.2594476.

Appendices

A Key Terminologies

- **Outlier Set**: A cluster of data points that collectively deviates significantly from the rest of the dataset, which we consider and evaluate as a single entity. Unlike traditional individual outliers, outlier sets represent groups of points that are anomalous together.
- **Inlier Set**: The main body of data points that represent typical or expected behaviour in the dataset. These points form the reference against which potential outlier sets are compared.

- **Floating Set**: In our synthetic datasets, this refers to a group of points deliberately placed at varying distances from the inlier set to test outlier detection. The position and spread of floating sets are controlled by parameters to evaluate method performance under different conditions.
- **Strict Ground Truth**: A floating set is labelled as an outlier set only if it has no points within the convex hull (boundary) of the inlier set. This represents the most conservative definition of what constitutes an outlier set.
- **Relaxed Ground Truth**: A floating set is labelled as an outlier set if its centroid (mean position) lies outside the convex hull of the inlier set, even if some individual points overlap with inliers. This allows for partial overlap between outlier and inlier sets.
- **Purity**: A metric measuring what percentage of points in a detected outlier set were actually generated as outliers in our synthetic datasets. Higher purity indicates more accurate outlier set detection.
- **Latin Hypercube Sampling (LHS):** A statistical method for generating near-random samples of parameter values from a multidimensional distribution, ensuring better coverage of the parameter space than pure random sampling.
- **Convex Hull:** The smallest convex set that contains all points in a dataset. In 2D, can be visualized as a rubber band stretched around the outermost points.
- **Hyperparameters:** Parameters that must be set before running the algorithm, such as the number of clusters K and weight threshold π_{th}.

B Determining The Number Of Clusters (K) For OSTI

In this study, we utilized Bayesian Information Criterion (BIC) (*Ishioka et al.*, 2005) to calculate optimal number of clusters. BIC helps select the optimal number of clusters among parametric models with different cluster counts and is represented by equation 8:

$$BIC = \ln(n) \cdot k - 2 \cdot \ln(L) \tag{8}$$

where:

- n is the number of observations;
- k is the number of clusters or parameters;
- L is the maximum likelihood of the model.

BIC penalises model complexity and is useful when trying to avoid overfitting. However, this assumes that your data follows a normal distribution, which may not always be true (*Cheong and Lee*, 2008; *Schwarz*, 1978). The optimal number of clusters is usually the one that minimises the criterion. Also, this study does not aim to present a systematic methodology for choosing K beyond the rule of thumb that it is appropriate.

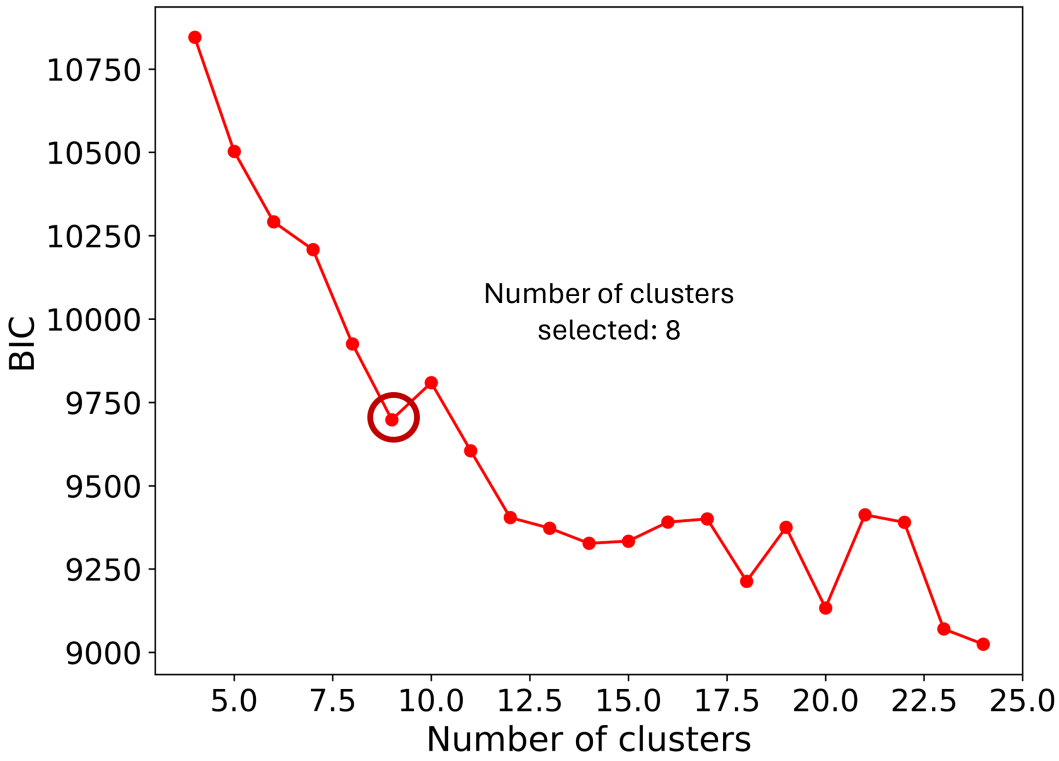

Figure 8: Number of clusters selection using Bayesian Information Criterion (BIC) applied to 2D data from Dolan Databases (DD). The x-axis shows the number of clusters (K) ranging from 1 to 20, while the y-axis represents the BIC score. The blue line shows how BIC values change as the number of clusters increases. The optimal number of clusters ($K = 8$) is identified at the point where adding more clusters provides diminishing returns in terms of model fit, indicated by the elbow in the BIC curve. This elbow point represents the balance between model complexity and goodness of fit - clusters beyond this point add computational complexity without substantially improving the model's ability to represent the data structure. The elbow (encircled) at $K = 8$ suggests an appropriate choice for OSTI's clustering step when analysing 2D irrigation withdrawal patterns.

The selection of $K = 8$ clusters is based on analysing the BIC curve behaviour across different cluster numbers as represented in Figure 8. While the first local minimum appears at $K = 8$, this selection is supported by several considerations:

1. **BIC Stability:** After $K = 8$, increasing K provides diminishing returns in terms of BIC improvement, suggesting that additional clusters are not capturing meaningful structure.
2. **Threshold Relationship:** For outlier detection with a weight threshold $\pi_{th} = 0.1$, $K = 8$ corresponds to an average cluster size of $\frac{1}{8} = 0.125$, slightly larger than the threshold. This allows smaller-than-average clusters to be identified as candidate outlier sets.
3. **Empirical Testing:** Through extensive testing on synthetic datasets with known ground truth, $K = 8$ consistently provided robust outlier set identification across different data configurations.
4. **Computational Efficiency:** While larger K values might marginally improve BIC, they increase computational cost without proportional gains in outlier detection accuracy.

C Synthetic Dataset Generation Using *make_ blobs*

The synthetic dataset (inliers and outliers) was generated using *make_ blobs*, a function from the Python library scikit-learn scikit-learn and is represented by the equation 9.

$$\begin{aligned} X, y = \text{make\_blobs}(&n_{\text{samples}}, n_{\text{features}}, \\ &\text{centers}, \text{cluster\_std}, \text{center\_box}, \\ &\text{shuffle}, \text{random\_state}) \end{aligned} \tag{9}$$

where,

- X is the generated sample features;
- y is the integer labels for the clusters;
- n_{samples} is the number of data points;
- n_{features} is the number of dimensions;
- centers is the number of centres to generate blobs;
- cluster_std is the standard deviation of the clusters;
- center_box is the bounding box to generate centres;
- shuffle whether to shuffle the samples;
- random_state is used for reproducibility.

Figure 9: Scatter plot displaying patterns for 1,000 synthetic datasets across Case 2 (two outlier sets). Each row shows a different inlier shape configuration (Circle, Ellipse, Triangle, Irregular) with four panels per shape: strict and relaxed conditions for outlier set 1 (Out1) and outlier set 2 (Out2). Each point represents a single dataset plotted by its distance (x-axis, 0-100 units) and angle (y-axis, 0-360 degrees). Points are coloured to show True Positives (TP, green), True Negatives (TN, blue), False Positives (FP, black), and False Negatives (FN, yellow), illustrating OSTI's detection performance across different geometric configurations and ground truth conditions.

D Case 2 Results

This section presents detailed visualization results for Case 2, where OSTI's performance is evaluated on detecting two outlier sets simultaneously. Figure 9 displays results across 1,000 synthetic datasets for each inlier shape configuration (Circle, Ellipse, Triangle, Irregular) under both strict and relaxed ground truth conditions. The plots track OSTI's detection performance for each outlier set (Outlier set 1 and Outlier set 2) separately to demonstrate the method's ability to detect multiple outlier sets within the same dataset.

Figure 9 illustrates these results through 16 panels, organized in four rows by shape. For the circle shape (panels a-d), detection patterns show clear separation between the two outlier sets, with Distance 1 vs Angle 1 plots revealing similar characteristics to Distance 2 vs Angle 2 plots. The ellipse configuration (panels e-h) maintains effective detection but shows more pronounced angular dependencies, particularly visible in the distribution of FN (yellow points) along the major axis. Triangle shape (panels i-l) demonstrates how OSTI adapts to sharp geometric features, with detection boundaries showing distinct shifts near vertex angles. The irregular shape (panels m-p) reveals OSTI's ability to handle complex boundaries while maintaining reliable detection of both outlier sets.

A detailed feature importance analysis for Case 2 through a heatmap visualization of parameter influence across different inlier shapes and ground truth conditions is represented in Figure 10. The analysis was conducted using random forest classifier information gain, examined three key parameters: distance 1, distance 2 (for OS1 and OS2 respectively), angle, and standard deviation for both outlier sets. For first outlier set (OS1), distance 1 consistently shows highest importance (0.720-0.759) across all shape configurations, particularly pronounced for irregular and circle shapes. Similarly, for second outlier set (OS2), distance 2 demonstrates dominant influence (0.670-0.777), with strongest effects in irregular and circle configurations. Angle and standard deviation parameters show relatively lower importance values (0.037-0.111), suggesting spatial positioning has greater impact on outlier set detection than cluster spread. This systematic analysis confirms that the radial distance of outlier sets from inlier centroids is the primary determinant of detection success, while angular position and cluster spread play secondary roles. This insight holds true across all geometric configurations and ground truth definitions.

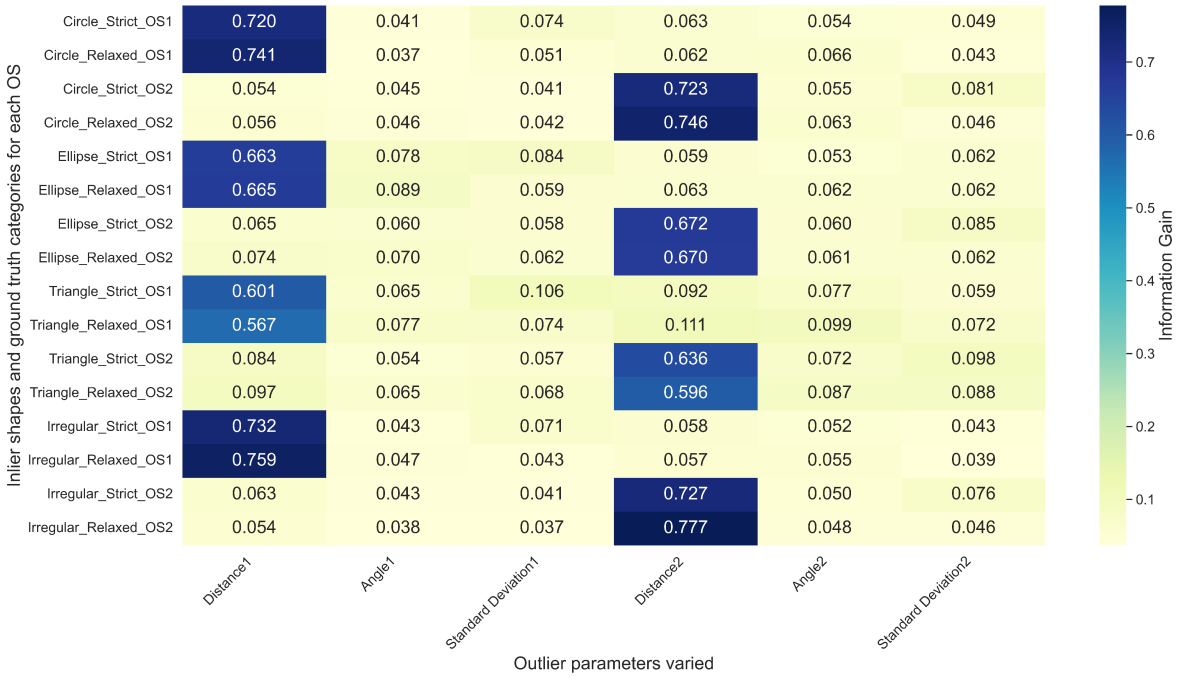

Figure 10: Heatmap showing the relative importance of each of the three parameters varied across Case 2 (Two outlier set) for both ground truth categories strict and relaxed for all four shapes (circle, ellipse, triangle, irregular) computed by information gain using random forest. Each cell in the heatmap represents the information gain of a parameter for a specific shape type and target ground truth condition. The colour intensity of the cells indicates the magnitude of the information gain, with darker colours corresponding to higher values. Here, distance 1 is the most influential parameter for outlier set 1 and similarly distance 2 is the most influential parameter for outlier set 2.