Environmental Modelling and Software 203 (2026) 107006

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft

Evaluating rule induction algorithms for scenario discovery

Jasper T. ter Horst, Patrick Steinmann *, Jan H. Kwakkel

Delft University of Technology, Delft, The Netherlands

ARTICLE INFO

Keywords:
Rule induction
PRIM
CART
Scenario discovery
Robust decision making
Deep uncertainty

ABSTRACT

Scenario discovery translates large simulation ensembles into interpretable input regions linked to policy-relevant outcomes. While previous studies have compared scenario discovery algorithms, they were ad hoc and hard to reproduce. We propose a general workflow to evaluate rule induction methods for scenario discovery. The workflow (i) provides synthetic benchmarks that expose axis and directional misalignment, nonlinearity, boundary fuzziness, and dimensional noise; (ii) unifies metrics and diagnostics around coverage–density trade-offs, interpretability, runtime, and scaling; and (iii) prescribes a staged experiment design from low-dimensional screening to stress testing. We illustrate the approach by comparing established algorithms PRIM and CART with an oblique decision tree variant called HHCART(D), finding that the latter does not outperform the former. Our workflow surfaces method-specific trade-offs and supports principled, reproducible algorithm selection for scenario discovery.

1. Introduction

In many environmental policy domains, decision-makers must act under conditions of *deep uncertainty*: the system structure is contested, key probabilities are unknown, and stakeholders disagree on which outcomes matter most (Lempert, 2003). To address these challenges, exploratory modelling has emerged as a way to map a wide variety of plausible futures and identify strategies that remain robust across them (Bankes, 1993; Lempert, 2003; Groves and Lempert, 2007). This approach has facilitated decision-support in a variety of socio-environmental domains such as climate adaptation, water and flood risk management, energy transitions, and tourism (Haasnoot et al., 2011; Kwakkel et al., 2013; Guivarch et al., 2016; Greeven et al., 2016; Student et al., 2020).

Scenario discovery is a key computational tool used to implement this exploratory approach (Schlumberger et al., 2026). It allows analysts to work backward from policy outcomes to deduce the combinations of input conditions that could cause them (Lempert et al., 2008). Using simulation models, analysts sample the uncertainty space and induce concise rules that summarise the input space into regions associated with policy-relevant behaviour (Lempert et al., 2008; Bryant and Lempert, 2010). This process directly addresses two long-standing challenges in environmental planning: how to systematically select a small number of scenarios from a vast space of possibilities, and how to meaningfully incorporate probabilistic or structural uncertainty (Lempert et al., 2008). The resulting concisely defined rules provide robust and adaptive planning guidelines and help surface critical vulnerabilities and opportunities without requiring consensus on how the world works or which single future is most likely (Groves and Lempert, 2007). We note here that, in this work, we use "rule induction" as a blanket term for identifying regions of input parameter space which generate certain model outputs of interest, although other authors varyingly prefer "rule induction", "learning", "subspace partitioning", "scenario partitioning", "classification", and other terms, sometimes depending on the specific algorithm employed.

Rule induction algorithms provide the primary operationalisation of scenario discovery because they identify explicit and interpretable conditions under which outcomes of interest occur. Prominent examples include the Patient Rule Induction Method (PRIM) (Friedman and Fisher, 1999; Lempert et al., 2008), Classification and Regression Trees (CART) (Breiman et al., 1984; Lempert et al., 2008), and rotated extensions such as PCA-PRIM (Dalal et al., 2013; Kwakkel et al., 2013). In a number of studies, analysts have assessed these algorithms in terms of how well they capture decision-relevant data points, how selective the identified scenarios remain, and how simply the resulting rules can be communicated in decision processes (Lempert et al., 2008; Bryant and Lempert, 2010; Dalal et al., 2013; Parker et al., 2015). However, these evaluations of scenario discovery methods have so far relied on ad-hoc test cases and experiment designs, limiting comparability across studies.

We introduce a workflow for evaluating the suitability of rule induction algorithms for scenario discovery. Our workflow proposes

* Corresponding author.
E-mail address: p.steinmann@tudelft.nl (P. Steinmann).

https://doi.org/10.1016/j.envsoft.2026.107006
Received 29 October 2025; Received in revised form 20 March 2026; Accepted 25 April 2026
Available online 6 May 2026

standardisation of three elements to enable consistent and reproducible comparisons across methods:

(i) A library of novel synthetic benchmarks designed to test explicit geometric challenges and augmented with controllable stressors like boundary fuzziness and dimensional noise.
(ii) A unified metrics and visualisation toolkit centred on the established criteria of scenario discovery (coverage, density, and interpretability) and complemented by diagnostics for runtime performance and scaling.
(iii) A reproducible experimental design that guides evaluation from initial screening on low-dimensional benchmarks to stress-testing against high-dimensional complexity.

We demonstrate the utility of this new workflow by using it to compare an algorithm from the Oblique Decision Trees (ODT) family – specifically HHCART(D) – with established rule induction algorithms. This demonstration illustrates how the workflow exposes trade-offs among coverage, density, and interpretability, and how it enables principled comparisons with established baselines such as PRIM and CART. Crucially, the workflow is method-agnostic and provides an extensible foundation for rigorous comparison of any rule induction approach for scenario discovery.

2. Background

2.1. Scenario discovery

Scenario Discovery is a key decision support tool for exploratory modelling and robust decision-making. It was developed in the early 2000s at the RAND Corporation as a structured, model-based alternative to narrative-based scenario planning (Groves and Lempert, 2007; Lempert et al., 2008; Bryant and Lempert, 2010). It was designed to address limitations in earlier approaches for decision-making under deep uncertainty (Lempert, 2003). Rather than constructing hypothetical futures *a priori* and observing their consequences, scenario discovery begins with a large ensemble of simulation-generated futures (sometimes referred to as Massive Scenario Generation (Davis et al., 2007)) and then systematically works backwards to identify the model conditions which are strongly associated with outcomes of interest. This reversal of logic allows analysts to pinpoint which uncertainties matter most for policy success or failure, without presupposing agreement on how the world works or which goals are normative. Different authors have classified scenario discovery as forms of sensitivity analysis (Pianosi et al., 2016) and vulnerability analysis (Bonham et al., 2025), respectively. Applications of scenario discovery include health systems (Götz et al., 2024), climate change adaptation (Guivarch et al., 2016), tourism (Student et al., 2020), logistics (Halim et al., 2016), and others. The conceptual idea of mapping certain classes of simulation model outputs to their generative input parameter ranges has also been picked up in other domains under various names and framings including behaviour analysis (Sücüllü and Yücel, 2014), pattern analysis (Edali, 2022), and simulation decomposition (Kozlova et al., 2024).

Technically, the process of scenario discovery relies on rule induction or machine learning algorithms to identify structured regions within a model's input space which are strongly associated with a subset of the model's outcomes. These regions provide compact, interpretable summaries of the conditions necessary to reach (or avoid, as the case may be) the type of outcomes included in the subset of interest. As Bryant and Lempert (2010) explain, the goal of scenario discovery is to "summarise sets of plausible future states of the world that illuminate key vulnerabilities [and opportunities] in proposed policies and to describe these scenarios in a manner useful for decision-makers and stakeholders".

The scenario discovery process typically includes three to four distinct steps:

1. **Sampling the Input Space.** The analysis begins by creating a simulation experimental design which systematically varies all uncertain input parameters of a simulation model. Because the model acts as a black box generator function (Lempert et al., 2006), any type of simulation model can be used, including system dynamics, agent-based modelling, and discrete event simulation.
2. **Identifying Outputs of Interest.** Among the simulation model's outputs, the outputs of interest (sometimes referred to as "decision-relevant" outputs (Bonham et al., 2025)) are identified. In the simplest case, this is done using a threshold criterion on a single output of the model (e.g., Greeven et al., 2016), classifying all model runs exceeding (or failing to reach, as the case may be) the threshold as being "of interest". However, more advanced approaches such as clustering model outcomes (Kwakkel et al., 2013; Steinmann et al., 2020) or multi-dimensional thresholds (Halim et al., 2016; Oostdijk et al., 2024) may also be used.
3. **Identifying parameter conditions.** A rule induction or machine learning algorithm is used to identify the model input parameter conditions (sometimes referred to as rule, box, region, or subspace) which are likely to generate a model output of interest. A number of algorithms have been proposed; we discuss them in the following subsections.
4. (Optional) Depending on the goals of the decision support process and the rule induction algorithm used, it may be desirable to either run the entire scenario discovery process multiple times for different outcomes of interest (Auping et al., 2015; Student et al., 2020) or repeat the last rule induction step to gain a more nuanced understanding of the model's input–output mapping (Guivarch et al., 2016; Lempert et al., 2008; Steinmann et al., 2024).

2.2. Algorithms for scenario discovery

A number of algorithms have been proposed for the third step of scenario discovery, including both decision trees and rule algorithms (Bénard et al., 2021). While these algorithms all broadly fulfil the intended purpose of scenario discovery, they use various approaches to perform the rule induction or classification, which may lead both to different decision support processes and outcomes. Fig. 1 shows three established approaches, which are discussed below.

The predominant algorithm used for scenario discovery is the Patient Rule Induction Method (PRIM, Friedman and Fisher, 1999). This is a rule-based algorithm (Bénard et al., 2021) designed to identify regions in a dataset where a particular variable has especially high values. The algorithm starts by considering the entire dataset (in the context of scenario discovery, this is the entire input parameter space of the model) and then iteratively peeling away small slices of data, depending on which slice removal most increases specific criteria related to the variable of interest. After the peeling process has found a subset of the data which maximises the variable of interest, a secondary pasting process is used to evaluate whether adding small slices of data back to this subset may further improve its representativeness. PRIM was adapted for scenario discovery using a binary classification (Lempert et al., 2008) of data points in the dataset, which allows for easier interpretation of the subset criteria because they can be expressed as coverage (total data points of interest included in the subset), density (ratio of interesting to uninteresting data points in the subset), and interpretability (number of peeled dimensions and boxes used). The most widely used software implementation of PRIM is part of the Python-based Exploratory Modelling & Analysis Workbench by Kwakkel (2017). An implementation in R is also available (Bryant, 2014).

A number of extensions and modifications for PRIM have been proposed. Kwakkel and Jaxa-Rozen (2016) implemented a modified

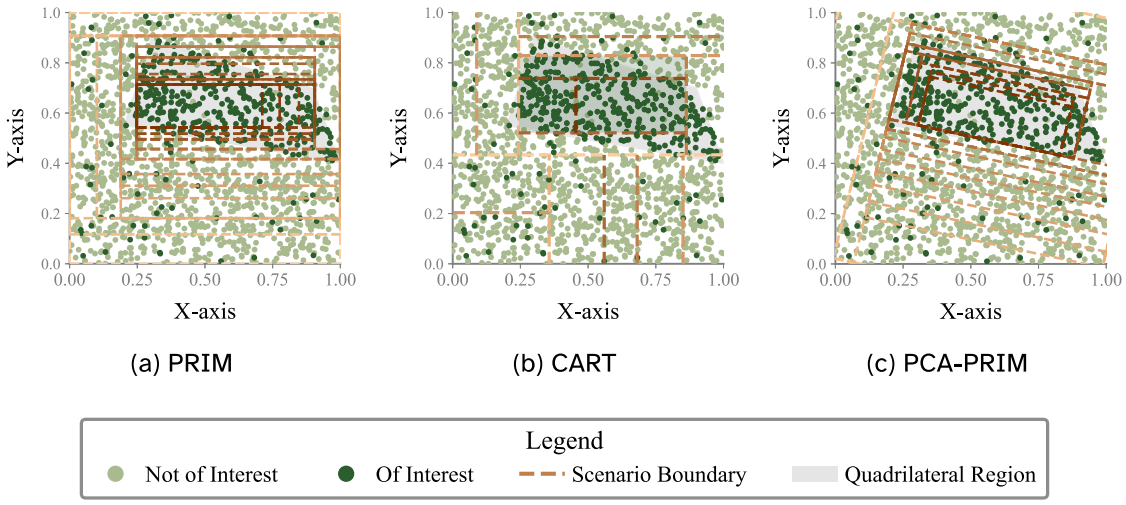


Fig. 1. A comparison of how PRIM, CART, and PCA-PRIM partition the same synthetic dataset. The dataset contains a rotated quadrilateral as the region of interest, designed to challenge axis-aligned methods.

objective function in PRIM which uses Gini impurity to achieve better peeling and pasting results. Dalal et al. (2013) used a Principal Component Analysis (PCA) pre-processing step to improve PRIM outcomes for outcome subsets of interest which were not orthogonally axis-aligned with the model's input parameter space by first applying a single global rotation to the model input parameter data. However, this does incur an interpretability penalty and struggles with categorical data. Rozenberg et al. (2014) explored an alternative approach for identifying multiple regions of interest, with the goal of finding more precise decision-relevant input parameter combination rules. Kwakkel and Cunningham (2016) used an ensemble of PRIM rule inductions performed on randomly permutated versions of a dataset to improve the quality of the decision-relevant rules. Finally, Kwakkel (2019) explored whether simultaneous many-objective optimisation of the PRIM criteria could yield more insightful results.

Classification and Regression Trees (CART, Breiman et al., 1984) are an alternative approach for rule induction in scenario discovery. CARTs, a form of decision trees (Bénard et al., 2021), are constructed by recursively applying binary, axis-aligned splits to the dataset, always selecting the split which immediately maximises a criterion, typically impurity. This process continues until a stopping criterion is met, often maximum depth or an impurity threshold. Lempert et al. (2008) explored CART for scenario discovery, highlighting that while CART has the benefit of being fully automated, it typically creates many splits, reducing interpretability, and maximises its criterion at each split, potentially sacrificing global pattern recognition for local purity optima. A number of authors have since employed CART for scenario discovery (e.g., Jafino and Kwakkel, 2021; Sunkara et al., 2023; Merino-Benítez et al., 2024), although the algorithm appears less widely used than PRIM.

2.3. Algorithm evaluations

We are aware of four distinct evaluations or comparisons of scenario discovery algorithms. Lempert et al. (2008) compared PRIM and CART, using four simple geometric shapes (tapered, intersecting barrels) in 3-dimensional and 10-dimensional space as test cases. van Droffelaar (2020) re-used these shapes to evaluate a novel rule induction algorithm called DREAM (Vrugt, 2016), which combines the first and third steps of scenario discovery into an adaptive sampling approach. van Droffelaar (2020) expanded the barrel shapes with additional dimensions (up to 50). Dalal et al. (2013) compared PRIM with PCA-PRIM using ten synthetic datasets, including eight 5-dimensional and two 15-dimensional problems. The datasets were intentionally not axis-aligned, as PCA-PRIM was designed to address the problem of axis misalignment in rule induction. Finally, Kwakkel and Jaxa-Rozen (2016) used two simulation-generated datasets as test cases to compare PRIM with an improved objective function with both CART and PRIM without the improved objective function. The two underlying simulation models were previously published by Toman and Lempert (2008) and Sassi et al. (2010), while the datasets generated with these models and used for the scenario discovery process were created by Bryant and Lempert (2010) and Rozenberg et al. (2014), respectively.

All aforementioned comparisons of scenario discovery algorithms used ad-hoc or bespoke test cases and metrics in their evaluations. This makes it difficult to compare the results of the individual evaluations. Furthermore, the results cannot readily be generalised, as they were produced using a limited set of test cases which were often designed with specific algorithms in mind. A principled workflow for comparing rule induction algorithms, as we propose here, would therefore be a worthwhile contribution to the literature and advance the state of the art in scenario discovery.

3. Evaluation workflow

In this section, we present an evaluation and benchmarking workflow for rule induction algorithms in scenario discovery. We first provide an overview of this iterative approach, followed by a discussion of relevant performance metrics and a set of algorithm-agnostic test shapes designed to evaluate algorithmic performance.

3.1. Overview

Fig. 2 illustrates the evaluation framework as a series of iterative phases designed to transition a rule induction method from theoretical promise to practical application. The process is initiated by the identification of a candidate algorithm hypothesised to address a specific limitation of established methods, such as the step-like effect created by axis-aligned splits (Lempert et al., 2008).

Before testing begins, the analyst must establish a unified measurement framework. This ensures that the candidate is evaluated on a commensurate basis with established baselines like PRIM and CART. By applying identical definitions of coverage and density across all methods, the workflow reveals objective performance trade-offs. Furthermore, this stage involves selecting appropriate proxies for interpretability to allow for a comparison of how easily the resulting scenarios can be communicated to stakeholders.

Once the metrics are defined, the candidate algorithm is tested in the benchmarking phase. In this stage, the analyst replaces complex simulation models with known geometric test shapes to transform the scenario discovery process into a controlled experiment. This substitution allows for the systematic application of the foundational pipeline detailed in Section 2.1: sampling the input space, identifying outputs of interest, and performing rule induction. By evaluating the candidate alongside established reference algorithms, this phase provides a direct, objective baseline for performance.

This comparison acts as the first critical decision point in the workflow: if the candidate cannot recover the geometric logic and statistical performance of these shapes as effectively as established methods, the process loops back to the identification phase for refinement. However, if the algorithm demonstrates sufficient geometric fidelity and favourable interpretability trade-offs, it proceeds to a more rigorous stress test where complexity is intentionally increased. By introducing dimensional noise, the analyst evaluates computational scaling and robustness — specifically, whether the algorithm can still identify the relevant signal amidst a high-dimensional space of uninformative variables. After a candidate proves its performance on these test datasets, it is deemed fit for comparison to real-world policy models.

3.2. Metrics

As described earlier, scenario discovery seeks to identify regions of a model's input parameter space that are strongly associated with decision-relevant or interesting model outputs. Lempert et al. (2008) proposed that these regions should satisfy three principal criteria:

1. *Coverage*: The regions should capture a high proportion of outcomes of interest in the dataset.
2. *Density*: The regions should contain a high proportion of outcomes of interest relative to all data points within the region.
3. *Interpretability*: The regions should remain simple to interpret and communicate for usage in policy contexts.

Scenario discovery inherently involves trade-offs between these three metrics. As Lempert et al. (2008) note, gains in one criterion often come at the expense of another. Expanding region boundaries to increase coverage typically lowers density, as more irrelevant data points are included. Conversely, more tightly defining regions to improve density may exclude relevant data points, thereby reducing coverage. Interpretability introduces a further constraint: achieving high coverage and density often necessitates more complex rule sets, such as multiple regions or more thresholded dimensions. These trade-offs imply that no single solution will simultaneously maximise coverage, density, and interpretability. Analysts therefore often examine a range of candidate solutions along the Pareto frontier maximising these metrics (Kwakkel, 2019) and apply expert judgement to select those most appropriate for the policy context (Bryant and Lempert, 2010). The value of scenario discovery lies in its ability to make these trade-offs explicit, enabling transparent and informed discussion of which uncertainties matter most and under what conditions policy-relevant outcomes arise.

3.2.1. Coverage

Coverage quantifies the extent to which a scenario region in the input space captures data points associated with a specified outcome of interest (Lempert et al., 2008; Bryant and Lempert, 2010). It measures the proportion of all interesting data points in the dataset that fall within the selected region(s). Thus, the metric evaluates how fully the identified rules represent the phenomenon of interest. A coverage value of 1 indicates that all relevant data points are captured; a value of 0 indicates that none are included. Coverage is therefore a criterion to be maximised.

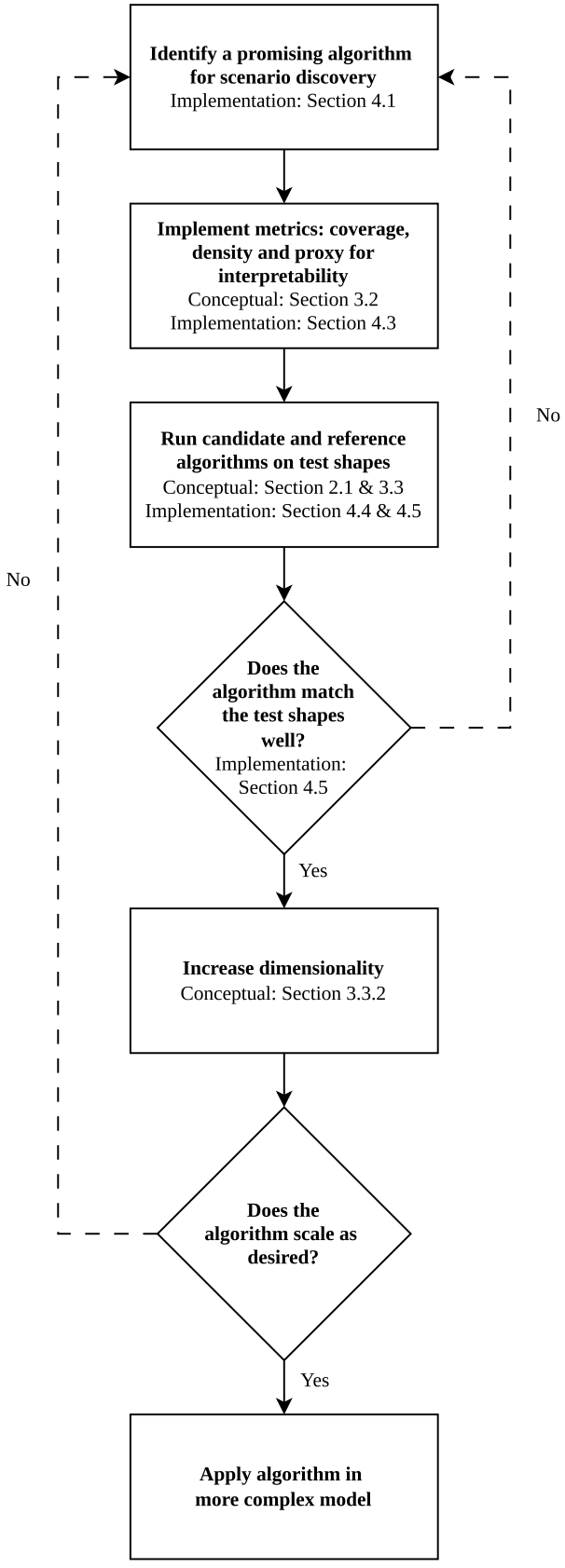


Fig. 2. Diagram of workflow for evaluating an algorithm's suitability for scenario discovery.

To formalise coverage, let X be the full set of data points, $X^I \subseteq X$ the subset of policy-relevant data points, and $R = \{R_1, \ldots, R_n\}$ the set of regions. Then, coverage is defined as:

$$\text{Coverage}(R; X^I) = \frac{\text{Number of data points of interest in the region}}{\text{Total number of data points of interest}} = \frac{\left|\{\, x \in X^I \mid x \in \bigcup_{i=1}^{n} R_i \,\}\right|}{|X^I|}. \quad (1)$$

3.2.2. Density

Density evaluates the selectivity of a scenario region by measuring the proportion of enclosed data points that correspond to the outcome of interest (Lempert et al., 2008, 2006). Decision-makers should find this measure important if they would like each scenario to be a strong predictor for the data points of interest (Bryant and Lempert, 2010). It is a precision-oriented metric: a value of 1 implies that nearly all data points within the region are outcome-relevant, whereas lower values indicate greater contamination by irrelevant data points. Density is therefore also a criterion to be maximised.

To formalise density, let X be the full set of data points, $X^I \subseteq X$ the subset of data points of interest, and $R = \{R_1, \ldots, R_n\}$ the set of regions. Then, density is defined as:

$$\text{Density}(R; X, X^I) = \frac{\text{Number of data points of interest in the region}}{\text{Total number of data points in the region}} = \frac{\left|\bigcup_{R_i \in R} x_j^I \;\middle|\; x_j^I \in R_i\right|}{\left|\bigcup_{R_i \in R} x_j \;\middle|\; x_j \in R_i\right|} \quad (2)$$

3.2.3. Interpretability

Interpretability describes the extent to which decision-makers can make sense of and apply the results of a scenario discovery analysis (Lempert et al., 2008). It is a prerequisite for policy relevance, as even statistically robust scenarios have limited value if they cannot be understood or meaningfully used in decision support processes. As Hadjimichael et al. (2024b) emphasise, outputs should not just represent uncertainty accurately, but also be cognitively accessible and practically actionable for their intended audiences.

Conceptually, interpretability is a criterion which should be maximised like coverage and density. However, there is no straightforward metric capturing interpretability, and indeed, there is little agreement on what even constitutes interpretability (Doshi-Velez and Kim, 2017; Lipton, 2018; Murdoch et al., 2019). Instead, suitable proxies must be identified based on the rule induction or machine learning algorithm in question (and, in applications, the decision support context). Below, we give an overview of some suggestions for metrics which proxy interpretability to some degree:

- Number of restricted dimensions. When using PRIM for scenario discovery, it is common to use the number of input parameter space dimensions for which rules are found as the proxy for interpretability (Lempert et al., 2008; Steinmann et al., 2024; Kwakkel, 2019). A lower number of restricted dimensions implies simpler and therefore more interpretable (Murdoch et al., 2019) results. This metric can be applied to any kind of rule-based algorithm.
- Tree depth. When CART is used with scenario discovery, the depth of the resulting regression tree may be used as a proxy for interpretability (Lempert et al., 2008; Birnbaum et al., 2022). Lower tree depth implies better interpretability; very deep trees may result in quite fine-grained analytical outcomes, or even overfit the underlying data. This metric is broadly applicable to decision tree-based methods.
- Sparsity of rules. For non-orthogonal (or "oblique") regression tree methods, the sparsity of the resulting trees is a key metric as these methods often draw upon all available input dimensions to identify optimal tree splits (Cañete-Sifuentes et al., 2021). This creates "dense" decision rules which involve dozens of model input parameter dimensions and are correspondingly difficult to interpret and translate into policy designs. Fewer dimensions included in every split imply better interpretability.
- Number of decision-relevant regions. A common approach for improving the accuracy of orthogonal rule induction approaches like PRIM or CART is to apply the algorithms iteratively, resulting in multiple regions in the input parameter space which are decision-relevant (Lempert et al., 2008; Gerst et al., 2013; Rozenberg et al., 2014; Guivarch et al., 2016). However, every additional region also reduces the overall interpretability of the analysis as the results become more fine-grained. A lower number of dimensions therefore implies better interpretability - Schwartz (1997) advises using no more than three or four regions, each structured around a sparse rule set covering two to three uncertainties. As both decision trees and rule-based algorithms identify decision-relevant regions, this metric could be used for both.
- Multi-metric approaches. When single metrics poorly approximate their target (as may be the case with the metrics listed above), using an ensemble may help explore trade-offs between them. For implementation or automation, aggregate metrics such as the f-score (Dalal et al., 2013) may be useful, although care should be taken to first explore how the metrics interact for different datasets (Manheim, 2023).

3.2.4. Runtime

As a decision support tool, scenario discovery is often used during stakeholder sessions to explore policy alternatives and their trade-offs. This requires that the used rule induction algorithm is reasonably fast, performing the requested analysis within seconds or perhaps tens of seconds to facilitate an ongoing dialogue and co-creation between the analyst, decision maker(s), and other affected parties (Führer et al., 2025). The runtime of an algorithm is therefore an important consideration. It is generally easy to measure (many programming environments have built-in support for timing code execution), but highly dependent on the used computing infrastructure, underlying data and algorithm settings. Important aspects of the data affecting runtime include the number of data points, the dimensionality of the dataset, and the type of data (binary, categorical, numerical, etc.). Beyond the conceptual and implementation aspects of the algorithms, settings which may affect runtime include the desired precision, search depth, and stopping conditions, which are often unique to the particular employed algorithm.

3.3. Synthetic test shapes

A rigorous evaluation of rule induction algorithms requires standardised benchmarks, yet the field of scenario discovery has largely relied on ad-hoc test cases. Seminal studies established the practice of using synthetic shapes to compare methods like PRIM and CART (Lempert et al., 2008; Dalal et al., 2013), but these bespoke cases have rarely been reused, preventing the emergence of a common standard for comparison. To address this gap and enable more comprehensive analysis, this paper introduces an expanded, open-library of synthetic test shapes where the ground truth is known *a priori*. This approach allows for a systematic assessment of an algorithm's ability to handle specific structural challenges.

The library comprises eight parametric shapes that define binary classification tasks in two and three dimensions (Fig. 3). As detailed in Table 1, these test shapes were specifically designed to probe the known limitations of established scenario discovery algorithms by incorporating three distinct geometric challenges:

Table 1
Geometric properties of the synthetic shapes used in this study. A checkmark indicates the presence of a property that is expected to challenge axis-aligned or globally rotated decision tree methods.

Shape	Axis misalignment	Directional misalignment	Nonlinearity	Likely to struggle
2D Rectangle	✓			PRIM, CART
2D Barbell	✓	✓	✓	(PCA-)PRIM, CART
2D Radial Segment	✓	✓	✓	(PCA-)PRIM, CART
2D Sine Wave	✓	✓	✓	(PCA-)PRIM, CART
2D Star	✓	✓		(PCA-)PRIM, CART
3D Barbell	✓	✓	✓	(PCA-)PRIM, CART
3D Radial segment	✓	✓	✓	(PCA-)PRIM, CART
3D Saddle	✓	✓	✓	(PCA-)PRIM, CART

(i) **Axis misalignment:** The decision boundary has a single, non-axis-aligned orientation, a property that challenges axis-aligned methods like PRIM and CART, which can only approximate such boundaries with a step-like series of orthogonal splits.
(ii) **Directional misalignment:** The boundaries exhibit multiple, locally varying *non-axis-aligned* orientations, a property that challenges both axis-aligned methods, which cannot efficiently capture any rotated feature, and globally-rotated methods like PCA-PRIM, which cannot use a single transformation to align with all relevant structural features simultaneously.
(iii) **Nonlinearity:** The decision regions are defined by curved boundaries, a property that fundamentally challenges all methods based on linear approximations. Both axis-aligned approaches like PRIM and CART, and globally-rotated approaches like PCA-PRIM, are structurally unable to capture such non-linear forms.

While more advanced shapes could be designed, Guivarch et al. (2016) showed that repeated application of scenario discovery essentially decomposes complex and/or unconnected shapes. We therefore limit ourselves here to simple, interpretable and visually inspectable patterns. If more elaborate test cases are desired, a simple approach would be to superimpose multiple (offset) basic shapes, as demonstrated by Lempert et al. (2008).

3.3.1. Boundary fuzziness

While the library of synthetic test shapes provides a rigorous foundation for assessing how algorithms handle specific geometric structures, these clean boundaries represent an idealised condition. In practice, a key challenge for scenario discovery is that the boundaries produced by complex, real-world models are rarely so sharp, a characteristic visually evident in applied studies (Gerst et al., 2013; Kwakkel, 2017). This phenomenon arises from several forms of model behaviour. First, many simulation models contain stochastic elements, meaning the same set of inputs can lead to different outcomes across separate runs. Second, highly non-linear relationships between inputs and outputs can create complex and unpredictable boundaries. Third, complex models are often highly sensitive to critical thresholds or tipping points, where a minor variation in an input can trigger a major shift in system behaviour. As a result of these effects, the boundary separating one class of outcomes from another is rarely a clean line, but is instead a zone where outcomes of interest and non-interest are mixed.

To ensure our benchmarks reflect this reality, we introduce a formal mechanism for emulating this zone of mixed outcomes. Boundary noise is added to each synthetic shape by probabilistically re-labelling points outside the true decision region. The likelihood of a point being re-labelled is designed to decay exponentially with its Euclidean distance from the true boundary, a rate governed by a parameter λ. As illustrated in Fig. 4, higher values of λ produce a wider and more ambiguous boundary zone, providing a more realistic test of an algorithm's ability to discern the true decision boundary amidst the noise inherent in complex system models.

3.3.2. Dimensions

The synthetic test shapes are further designed to address the challenge of high-dimensional uncertainty spaces. Policy models are frequently characterised by many parameters, yet a common insight from Global Sensitivity Analysis (GSA) is the principle of sparsity of effects: the determinative behaviour of a complex system is often driven by a small subset of influential parameters and their interactions (Box et al., 2005). An effective scenario discovery algorithm must therefore be capable of identifying this sparse set of drivers while remaining robust to the presence of many uninformative variables.

The proposed workflow tests this capability by augmenting datasets with dimensional noise; statistically independent, uniformly random features are added to create input spaces of up to 30 dimensions. This approach follows the precedent of earlier benchmarking studies that have explored the impact of increasing dimensionality (Dalal et al., 2013; van Droffelaar, 2020) and allows for a rigorous evaluation of whether an algorithm can effectively distinguish the signal (the sparse, influential factors) from the noise in a high-dimensional context. Adding dimensions also helps explore how a given algorithm scales in terms of runtime as the dimensionality of the underlying dataset increases.

3.4. Visualisations

Visual diagnostics are a cornerstone of our proposed evaluation workflow, offering a suite of tools that enable analysts to move beyond aggregate metrics for a more nuanced assessment of algorithmic behaviour. The workflow employs a range of visualisations, some of which are generic to any rule induction method, while others are specifically designed to probe the characteristics of the algorithms under review, such as the tree-based methods demonstrated in Section 4 of this paper. Rather than a prescribed sequence, these visualisations form a toolkit, allowing an analyst to select the appropriate lens for a given evaluation task.

A primary category of tools addresses the crucial criteria of interpretability and structural complexity, which are often in tension with raw statistical performance. Central to this is the coverage-density trade-off plot (Fig. 5(b)), which provides a holistic view of an algorithm's output by mapping coverage and density and annotating each point with an interpretability proxy, such as the number of decision-relevant regions. To complement this holistic view with a qualitative check, the workflow includes the decision boundary overlay plot (Fig. 5(a)). For low-dimensional benchmarks, this visualisation draws an algorithm's splits directly onto the data, offering an immediate, intuitive assessment of its geometric strategy; it makes issues like fragmentation, misalignment, or inefficiency visually obvious, providing crucial context that helps explain statistical performance.

To further assess interpretability, our workflow emphasises and includes tools for the direct inspection of a model's internal logic and rule complexity. For tree-based methods, for example, tree structure diagrams (Fig. 5(c)) offer a transparent view of the model. They reveal the exact rules at each node, making the overall depth and the specific algebraic complexity of the splits immediately apparent. This provides

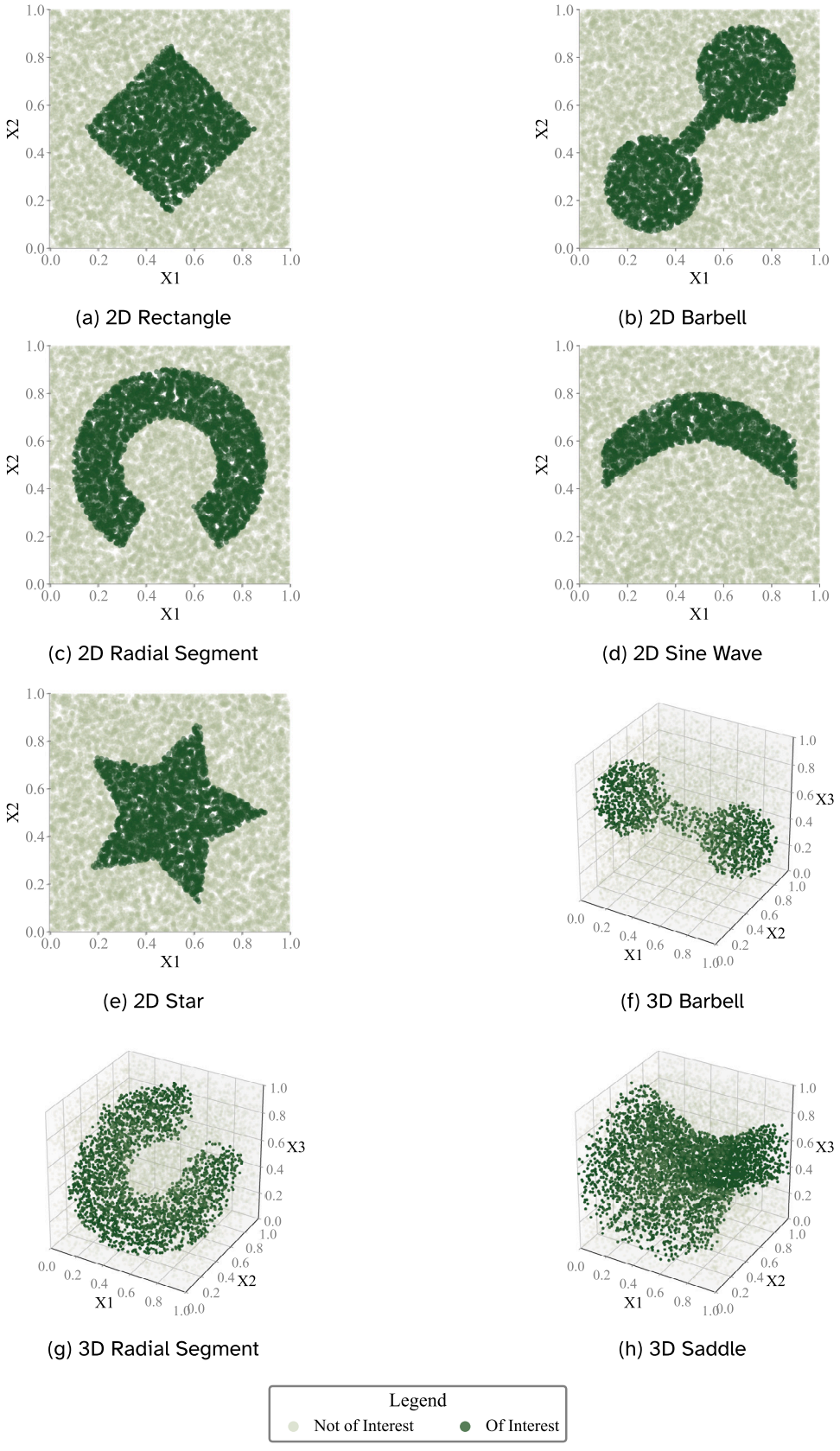


Fig. 3. The eight synthetic shapes used for benchmarking. Panels (a)–(e) show the 2D shapes, and panels (f)–(h) show the 3D shapes. A shared legend is provided at the bottom to reduce visual clutter.

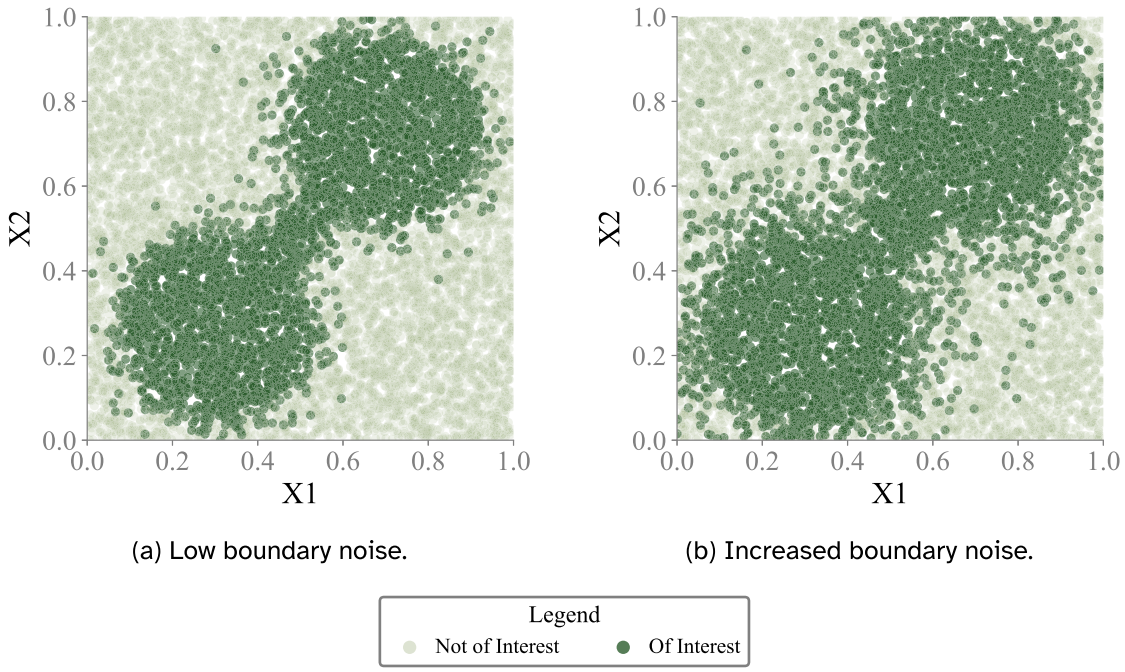


Fig. 4. Visual effect of boundary fuzziness on the 2D barbell shape. Panel (a) shows low boundary noise, where only a few points near the decision boundary are relabelled ($\lambda = 0.03$). Panel (b) shows higher boundary noise, resulting in a wider boundary zone ($\lambda = 0.07$).

the ground truth for the scenario's logic, which must ultimately be communicated to stakeholders. This qualitative inspection is complemented by sparsity plots (Fig. 6), which are particularly crucial for non-axis-aligned decision trees that risk producing overly complex rules. By tracking the average number of active features used per split, this plot quantitatively measures rule complexity. A low feature count indicates that the algorithm is finding sparse, simple rules, a highly desirable characteristic for generating actionable insights from high-dimensional policy models.

The second category of tools focuses on the quantitative assessment of performance, robustness, and efficiency. Performance trajectory plots (Fig. 7) are a primary tool for comparing algorithmic efficiency. They plot core metrics—coverage, density, and runtime—against a proxy for model interpretability, such as tree depth or the number of identified regions, allowing for a direct comparison of how effectively different algorithms achieve high performance with simple, interpretable models. This analysis is extended by robustness trajectory plots (Fig. 8), which stress-test an algorithm by comparing its performance trajectories under varying conditions, such as increasing dimensional noise, boundary noise, or different sample sizes. Finally, computational scaling plots (Fig. 9) diagnose an algorithm's feasibility for large-scale problems by revealing the computational complexity (log–log plots of runtime versus feature count or sample size), providing critical insight into whether an algorithm will remain practical for real-world policy models.

4. Demonstration

To demonstrate the utility of our evaluation workflow, we apply it to conduct a systematic evaluation of a theoretically promising but untested class of rule induction methods. Following the operational logic defined in Section 3.1, this demonstration executes the workflow by identifying a candidate algorithm, defining a commensurate set of metrics, and performing a comparative analysis against established baselines.

4.1. Candidate identification: Oblique decision trees

Building on the structural limitations discussed in Section 2.1, we identify Oblique Decision Trees (ODTs) as a candidate for overcoming the constraints of axis-parallel and globally rotated induction. While established decision tree methods such as CART are restricted to hyper-rectangular partitions, ODTs provide a theoretical alternative by utilising hyperplanes to partition the input space at any orientation. Unlike their axis-aligned counterparts, ODTs define decision boundaries as a linear combination of multiple input features:

$$\mathbf{w}^{\top}\mathbf{x} \leq v, \tag{3}$$

where $\mathbf{w} \in \mathbb{R}^d$ is a weight vector, $\mathbf{x}$ is the input vector, and v is a threshold (or bias coefficient).

We expect that this fundamental shift in split geometry directly addresses the axis misalignment and directional misalignment stressors identified in Table 1. By allowing hyperplanes to orient themselves according to the structure of the subspaces rather than the coordinate axes, ODTs theoretically resolve axis misalignment more efficiently than axis-parallel methods. This expectation is supported by the broader classification literature, which indicates that ODTs often achieve higher accuracy with fewer nodes and shallower depth than axis-aligned trees (Murthy et al., 1994; Wickramarachchi et al., 2016). Furthermore, because the optimal orientation is determined independently at each node, ODTs offer a "local adaptivity" that addresses directional misalignment — the challenge of scenario borders that shift orientation across different regions of the input space. This represents a potential advantage over globally-rotated methods like PCA-PRIM, which are restricted to a single transformation of the feature space and cannot adapt to locally varying structural orientations (Dalal et al., 2013).

Whether this flexibility also improves performance for nonlinearity is a central question for our benchmark execution. Although ODTs are more geometrically expressive than CART and PRIM, their splits remain fundamentally linear hyperplanes that cannot naturally "curve" to follow a boundary. We expect potential struggles with concave shapes like the 2D Barbell, where the algorithm may be forced to split through the "hollow" inner borders of the distribution rather than aligning with

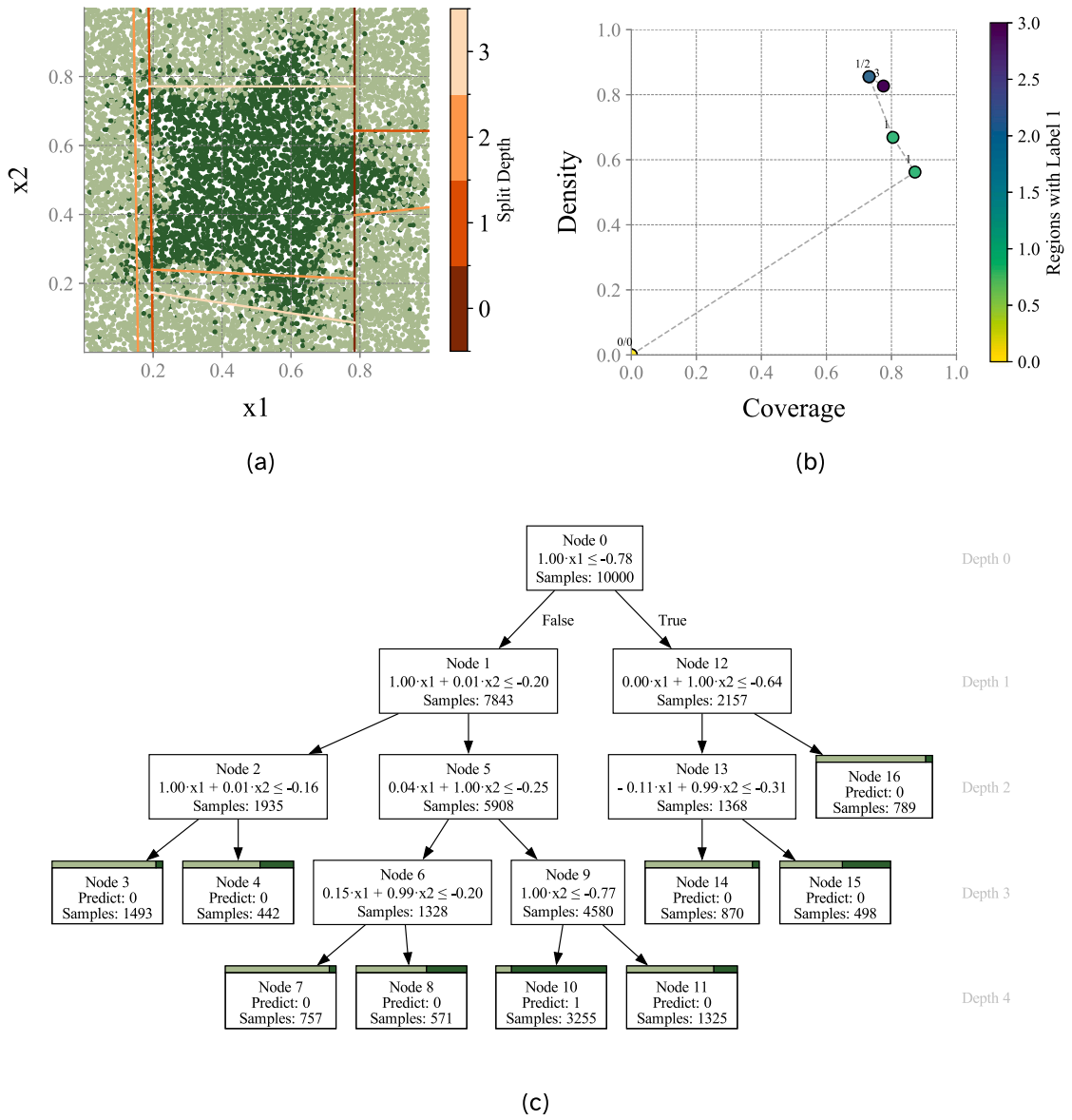


Fig. 5. A suite of visualisations for evaluating a tree-based algorithm (HHCART(D), a CART variant described in detail in Section 4.2). (a) The decision boundary overlay shows splits on the data, coloured by depth. (b) The coverage-density trade-off plot shows performance at each tree depth, annotated by the number of decision-relevant regions. (c) The tree structure diagram reveals the specific rules at each node.

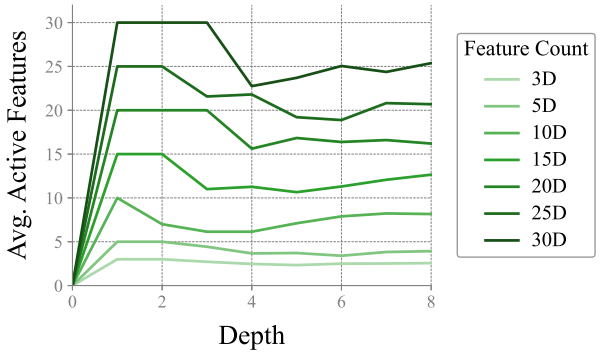


Fig. 6. Example of a sparsity plot, showing the average number of active features per split for HHCART(D).

them. In such cases, the efficacy of the ODT is expected to hinge on whether its local adaptivity can effectively approximate these complex boundaries through the strategic use of multiple linear subspaces.

Finally, we anticipate that any potential for improved geometric performance is met by the critical challenge of interpretability. Unlike a simple single-variable threshold, a linear combination (for example, $0.6 \cdot \text{GDP Growth} - 0.4 \cdot \text{Fuel Price} + 1.2 \cdot \text{Investment Risk} \leq 2.5$) is significantly more difficult for stakeholders to validate or communicate. We therefore test for a significant trade-off: while ODTs may produce scenarios with higher coverage and density, the resulting rules may be harder to translate into actionable policy insights than those produced by PRIM or CART.

4.2. Algorithmic implementation: HHCART(D)

The selection of a specific algorithm to represent the ODT family followed a preliminary screening of six candidates representing analytical, iterative, and gradient-based split-induction strategies. As

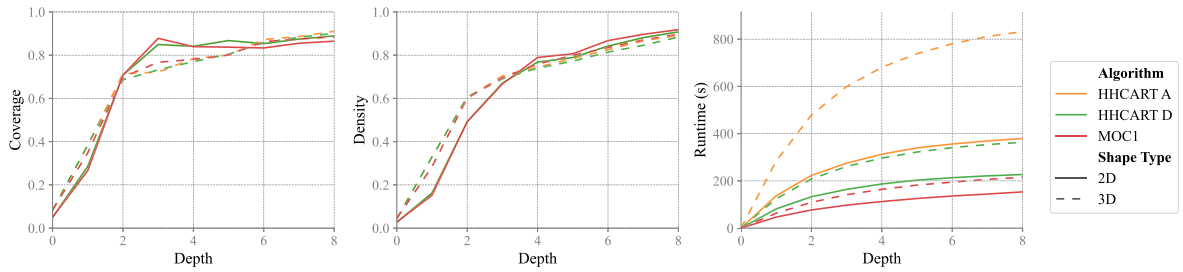


Fig. 7. Example of performance trajectory plots, showing coverage, density, and runtime against tree depth for three algorithms.

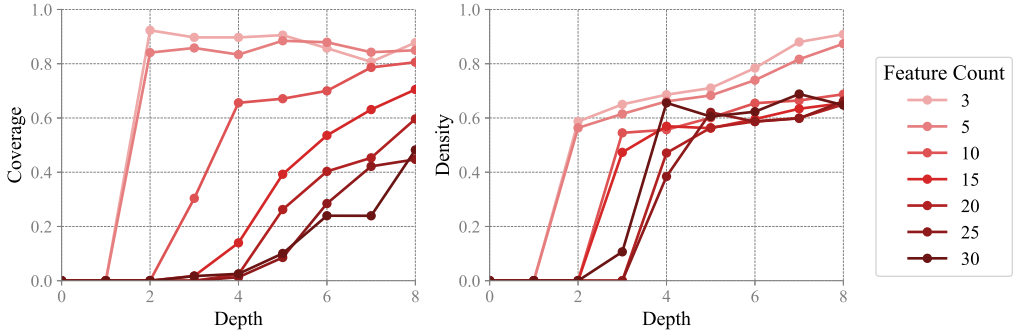
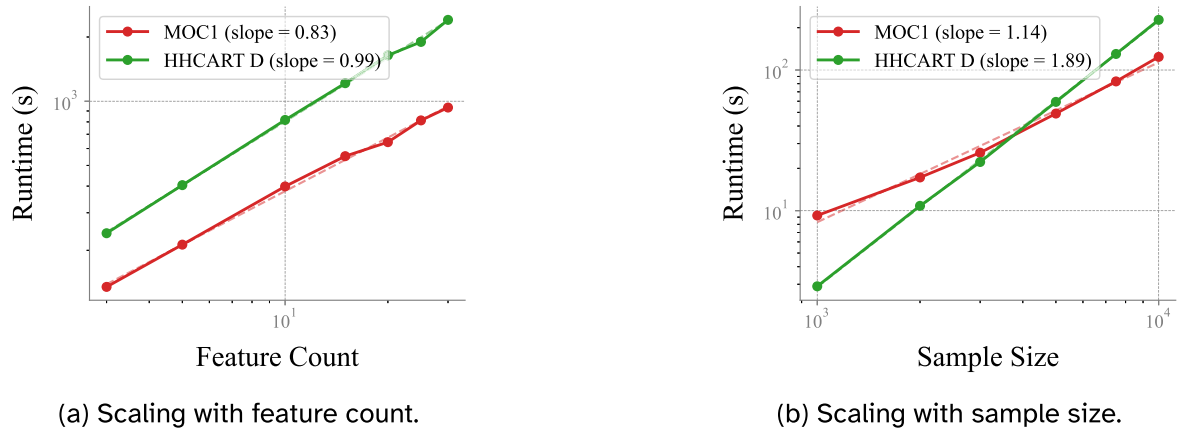


Fig. 8. Example of a robustness trajectory plot, showing the performance of the MOC1 algorithm under increasing dimensional noise.



(a) Scaling with feature count.

(b) Scaling with sample size.

Fig. 9. Log–log plots showing runtime scaling for two different rule induction algorithms (MOC1 and HHCART(D)). Dashed lines indicate fitted power-law trends. Panel (a) varies dimensional noise, while Panel (b) varies sample size.

described by ter Horst (2025), this screening process was used to identify which candidate algorithms best serves the requirements defined in Section 4.3 – specifically the need for high coverage and density at shallow tree depths, as interpretability was expected to be quite similar across oblique trees. Computational tractability and resistance to dimensional noise were also tested in this evaluation. Based on this evaluation, HHCART(D) (Wickramarachchi et al., 2016) was selected as the candidate algorithm for this demonstration.

HHCART(D) builds upon the standard CART framework but incorporates Householder reflections to determine split orientations. At each decision node, the algorithm identifies the dominant axes of variation within the data by performing an eigendecomposition on class-specific covariance matrices. It then calculates a reflection vector based on the dominant eigenvector from each class, which defines a transformation mapping the data into a temporary coordinate system.

In this reflected space, the algorithm identifies the optimal axis-parallel split, which is then projected back into the original feature space as an oblique hyperplane. To prevent unnecessary complexity, HHCART(D) compares this oblique split against the best axis-aligned alternative at every node, selecting whichever yields the greater reduction in impurity. This dual-search strategy should allow the model to remain axis-parallel where the data structure permits, while utilising oblique hyperplanes to resolve the more complex misalignment stressors discussed in Section 4.1.

4.3. Implementation of metrics

To evaluate the ODT candidate against established baselines, we operationalised the principal criteria for scenario discovery – coverage, density, and interpretability (Lempert et al., 2008) – using an appropriate set of metrics. Because scenario discovery algorithms often employ fundamentally different induction logics, establishing a common measurement framework is essential for a fair comparison. For the tree-based algorithms (CART and HHCART(D)), coverage and density were calculated over the union of all terminal nodes classified as the outcome of interest, allowing their performance to be compared directly against the single-box or multi-box results produced by PRIM.

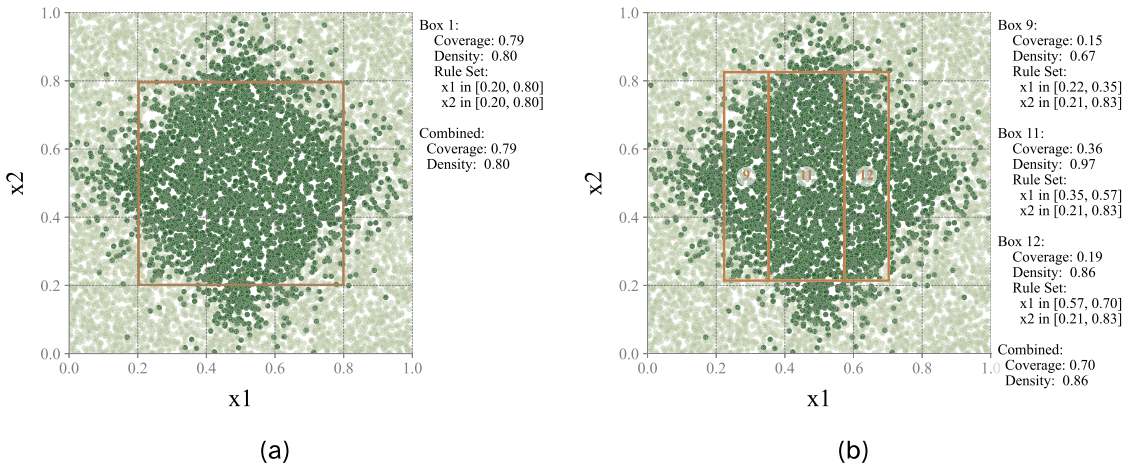


Fig. 10. Results of PRIM and CART on the rotated 2-dimensional rectangle benchmark under moderate boundary noise ($\lambda = 0.05$). Panel (a) shows the axis-aligned box selected using PRIM (*minimum mass* = 0.05, *density threshold* = 0.8). Panel (b) shows the CART result (*minimum mass* = 0.10).

Assessing interpretability required the use of specific quantitative proxies. The primary proxy was the number of regions classified as class 1, which directly measures the fragmentation of the final scenario. This is a crucial metric for tree-based methods like CART and ODT algorithms, which automatically generate a set of positive regions, in contrast to PRIM and PCA-PRIM, where the analyst manually selects a number of scenario boxes. A lower number of these positive regions indicates a more coherent, and therefore more interpretable, result. The second proxy was the algebraic complexity of the decision rules, allowing for a direct comparison between the simple, single-variable axis-aligned splits produced by PRIM and CART, and the more complex, multi-variable oblique splits generated by ODT algorithms.

4.4. Experimental design

To provide a rigorous test of the practical utility of Oblique Decision Trees, we designed a direct comparative experiment comparing HHCART(D) with PRIM and CART on simple benchmark shapes. We performed the comparison on two two-dimensional benchmark shapes, namely the rotated rectangle and the barbell. The two-dimensional nature of these problems was chosen specifically to allow for direct visual inspection of the resulting scenario partitions. These shapes present distinct geometric challenges: a simple, convex rotated structure and a fragmented, non-convex region. For each shape, 10 000 data points were generated using Latin Hypercube Sampling. A moderate boundary noise of $\lambda = 0.05$ was applied to simulate the ambiguity inherent in real-world simulation models.

To ensure a meaningful comparison, the parameters for each algorithm were selected to find a favourable balance between coverage, density, and interpretability, while also limiting the over-segmentation highlighted by Lempert et al. (2008). For PRIM, a peeling fraction of 5% and a minimum mass of 5% were used. For the tree-based methods, the primary lever to prevent over-fragmentation was the minimum mass for a terminal node. In the case of HHCART(D), this was complemented by a minimum purity constraint. Both parameters were tuned to produce more interpretable partitions, accepting a potential trade-off in lower density to avoid creating fragmented regions that would require manual recombination. The final PRIM scenario boxes were selected manually from the peeling trajectories to optimise the trade-off between coverage and density.

The baseline algorithms were implemented using the Exploratory Modelling and Analysis Workbench (Kwakkel, 2017), while HHCART(D) was implemented as a custom Python class adapted from the work of Majumder (2020). To ensure a consistent visual comparison across all methods, we developed a number of bespoke visualisations highlighted in Section 3.4, which we used to generate all scenario partition plots presented in this paper.

4.5. Analysis of results

The performance of each algorithm on the two benchmark problems reveals distinct trade-offs between geometric flexibility, statistical performance, and interpretability.

4.5.1. Rotated rectangle benchmark

On the rotated rectangle benchmark, PRIM produced an outcome that serves as a clear baseline for an interpretable, single-region solution. As shown in Fig. 10(a), the result is a single, axis-aligned box achieving 79% coverage and 80% density. This performance is inherently constrained by the method's geometry; its axis-aligned orientation cannot match the 45-degree rotation of the target region, leading to a visible mismatch at the corners and edges. It is important to note that this specific box represents just one point on the coverage-density peeling curve, which offers the analyst a range of candidate solutions.

Consistent with the observation by Lempert et al. (2008) that the CART algorithm "may proliferate the number of boxes", its solution for this benchmark was structurally more complex, requiring three separate boxes to describe the region (Fig. 10(b)). The combination of these boxes resulted in a direct performance trade-off against the selected PRIM box: a higher density of 86% was achieved at the cost of a lower coverage of 70%. Crucially, CART's automated, greedy procedure offers no flexibility in this outcome, presenting a single, fixed result. Given its higher initial complexity and the lack of choice for the modeller, the CART solution provides no compelling advantage over PRIM for this benchmark.

HHCART(D), in turn, failed to capitalise on its geometric flexibility, creating a result that did not effectively match the rectangle's simple structure (Fig. 11(a)). The final model presented is a regularised tree of depth four, chosen for its favourable performance on the coverage-density curve (Fig. 11(b)). At this depth, where further splitting was halted by the applied minimum mass and purity constraints, the model achieved a final coverage of 80% and a density of 84%. This performance, however, offered only a marginal improvement in the coverage-density trade-off compared to the simpler results of PRIM and CART.

Beyond its mediocre statistical performance, the partition generated by HHCART(D) was also the least interpretable. An inspection of the tree structure (Fig. 11(c)) reveals a complex and counter-intuitive description: the identified scenario is fragmented into two separate regions, and the algorithm's heuristic failed to find the four simple oblique lines that would have defined the target shape. This outcome is a direct consequence of the algorithm's greedy, node-by-node optimisation and constrained split-search strategy.

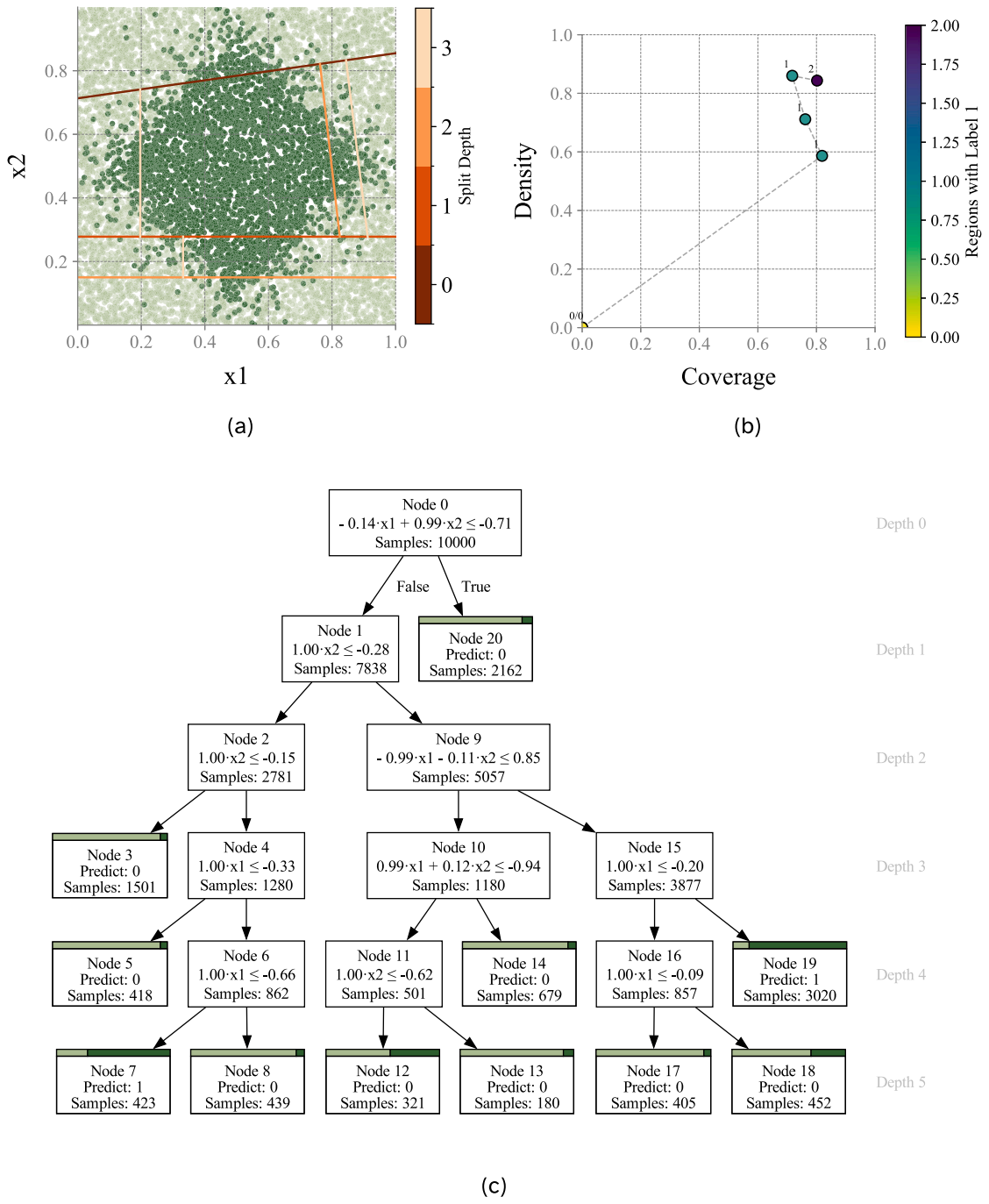


Fig. 11. HHCART(D) on the 2-dimensional rectangle benchmark under moderate boundary noise ($\lambda = 0.05$). Panel (a) overlays the oblique decision boundaries up to depth 4. Panel (b) shows the coverage-density trajectory across tree depths and number of regions classified as 1, up to depth 8. Panel (c) visualises the corresponding tree structure up to depth 4.

4.5.2. Barbell benchmark

On the barbell benchmark, PRIM again provided a simple and highly interpretable outcome, identifying two axis-aligned regions centred on the two circular regions (Fig. 12(a)). The selected boxes covered 68% of the target region, with a high density of 93%. Notably, no additional box was selected for the low-density connecting piece, as no candidate in the peeling trajectory satisfied the minimum mass and density thresholds. The final result thus offered a compact and transparent description of the two main regions of interest, prioritising high density over capturing the entire shape.

In contrast, CART achieved higher coverage than PRIM, but at the cost of increased fragmentation and reduced density (Fig. 12(b)). It used four axis-aligned boxes to approximate the two circular regions and portions of the connecting piece, reaching a combined coverage of 77% with a density of 90%. By including more of the mixed-class region surrounding the barbell, CART improved coverage but lowered density and increased model complexity, offering no clear performance advantage over PRIM's more focused and interpretable result.

For HHCART(D), we selected the model at search depth 4 for detailed analysis. At this stage, it captures the entire barbell in a single, large oblique space, achieving a high coverage of 88% with a modest density of 79% (Fig. 13(a)). This performance does not represent a clear improvement over the other methods, but rather a specific trade-off:

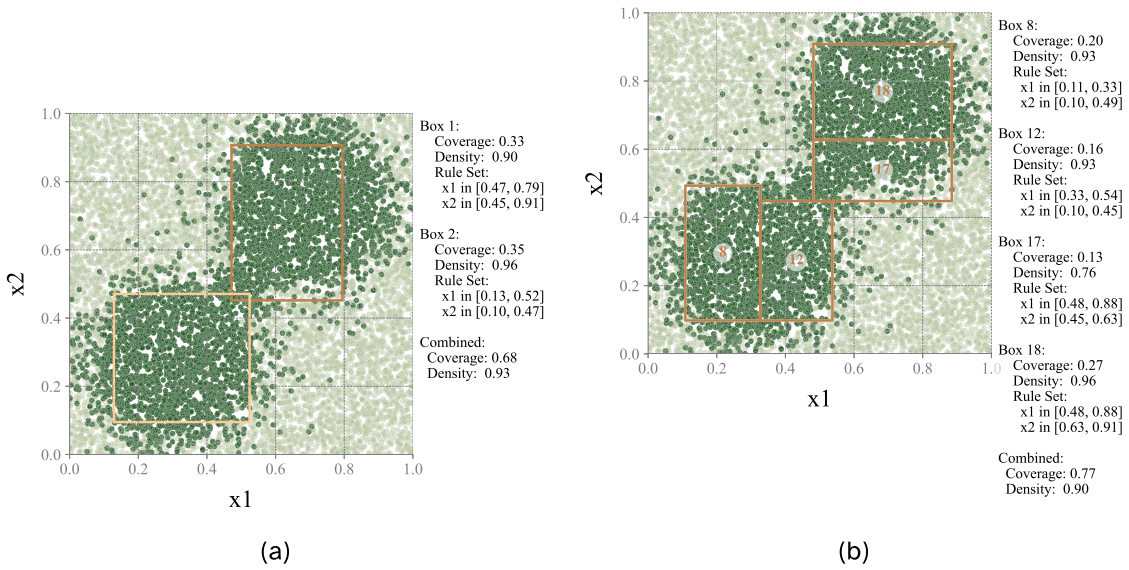


Fig. 12. Results of PRIM and CART on the rotated 2-dimensional barbell benchmark under moderate boundary noise ($\lambda = 0.05$). Panel (a) shows the axis-aligned box selected using PRIM (*minimum mass* = 0.05, *density threshold* = 0.8). Panel (b) shows the CART result (*minimum mass* = 0.07).

density is sacrificed to gain coverage by including the noisy, mixed-class regions adjacent to the barbell's connecting piece, which are not part of the core regions of interest. Furthermore, this single-region solution is already complex from an interpretability standpoint; the tree structure (Fig. 13(c)) shows it is defined by four separate oblique inequalities, each a linear combination of the input variables.

The coverage-density trajectory (Fig. 13(b)) illustrates the alternative trade-offs available beyond the depth-four model. For instance, at depth six, the model achieves a balanced state of 83% coverage and 83% density. While these statistics are favourable compared to those generated with PRIM, they come at a severe cost to interpretability. At this depth, the single region is fractured into three separate regions, which are collectively defined by seven complex oblique inequalities. This demonstrates the algorithm's unfavourable trade-off: to achieve a statistical performance similar to a simple method like PRIM, HHCART(D) produces a scenario that is substantially more fragmented and less interpretable.

4.6. Interpretation

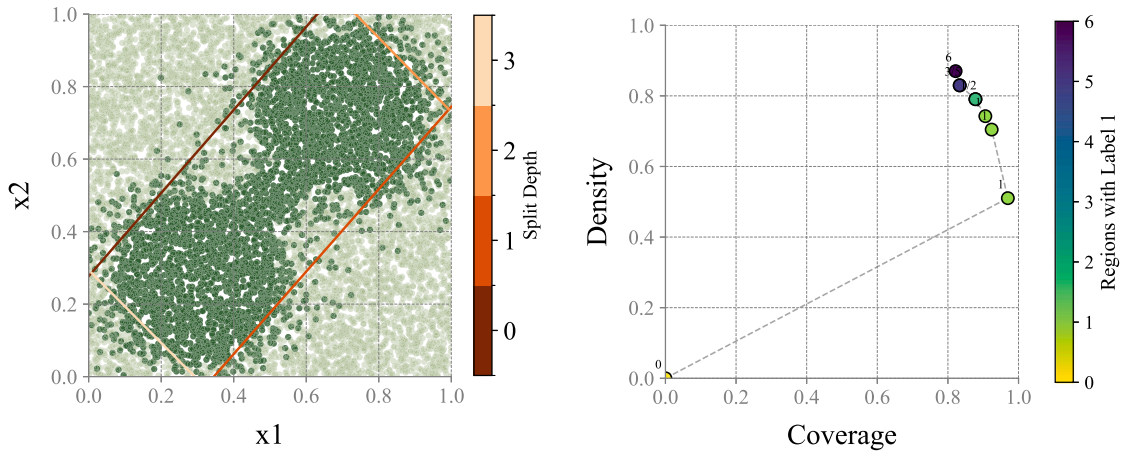
The results on the benchmark tests highlight the distinct trade-offs offered by each method. PRIM consistently served as the benchmark for interpretability, producing simple, high-density scenarios. A key feature of its approach is the analyst's ability to select a final box from a peeling trajectory, allowing for a deliberate choice in the coverage-density trade-off. In comparison, CART's automated procedure produced a single, fixed outcome that was typically more fragmented. While its statistical performance varied relative to the chosen PRIM box – sometimes offering higher density, other times higher coverage – it never demonstrated a compelling, decisive advantage that would justify its lower interpretability and lack of user flexibility. Compared to the established methods, HHCART(D) consistently yielded the most complex and least interpretable scenarios, forcing an unfavourable trade-off where any marginal gains in statistical performance came at a great cost to interpretability.

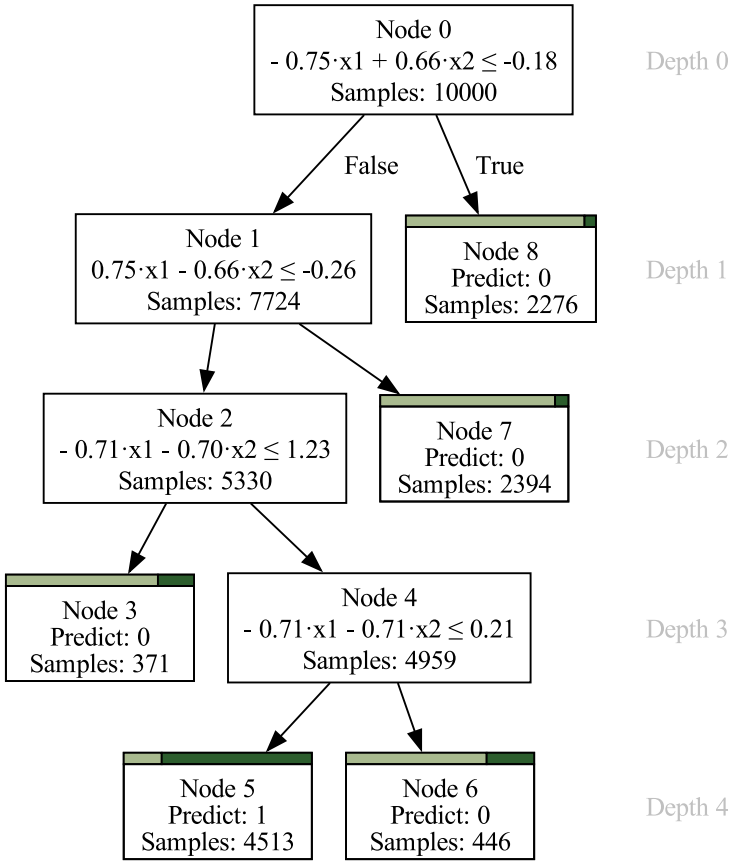
The underperformance of HHCART(D) can be attributed to two compounding weaknesses in its design: its greedy, node-by-node construction and its constrained split-search heuristic. First, like all CART-based algorithms, it is greedy, selecting the single split at each node that provides the greatest immediate reduction in impurity, without a global strategy for how multiple splits might combine to form the most efficient tree. Second, the search for this locally optimal split is itself limited by the strategy used. The algorithm uses a heuristic that evaluates a set of pre-determined candidate directions. If the true decision boundary does not align with one of these directions, the algorithm is structurally incapable of finding it. This combination can force the algorithm to select suboptimal splits, and consequently, it must grow a deep and convoluted tree to achieve high performance. This creates a trade-off where high statistical performance can only be reached at the cost of low interpretability.

This outcome demonstrates the methodological value of the proposed evaluation framework. The use of simple, low-dimensional benchmarks serves as a powerful and efficient screening mechanism. The visual inspection of the results is sufficient to conclude that HHCART(D)'s theoretical flexibility does not translate into a practical advantage over established methods in this context. This provides a principled justification for aborting further investigation (as shown in Fig. 2), thereby preventing the investment of significant computational resources and time into testing an unsuitable algorithm on more complex, high-dimensional problems where its deficiencies would be less transparent.

5. Discussion

Scenario Discovery is a powerful analytic method for exploratory modelling and decision support and has been applied in a wide variety of environmental policy domains. Given the academic and policy interest in the method, it is not surprising that a number of modifications and improvements have been proposed. However, the evaluations of these improvements have been ad hoc and case-based, limiting our ability to generalise insights. Our workflow provides a principled approach to evaluating potential algorithmic improvements to Scenario Discovery. It includes custom synthetic shape generators and is the first scenario discovery toolkit to incorporate configurable boundary noise or fuzziness, which reflects the complexity of large policy-relevant simulation models. In addition to standard scenario discovery metrics (coverage, density), our workflow also captures structural characteristics such as tree depth, number of regions, and feature usage per split, and provides rich visualisation tools, including tree diagrams, decision boundaries, and coverage–density trajectories, that support both diagnostic analysis and interpretability assessment. Together, these features establish a practical, extensible standard for future scenario discovery algorithm development and evaluation.





(c)

Fig. 13. HHCART(D) on the 2-dimensional barbell benchmark under moderate boundary noise ($\lambda = 0.05$). Panel (a) overlays the oblique decision boundaries up to depth 4. Panel (b) shows the coverage-density trajectory across tree depths and number of regions classified as 1, up to depth 8. Panel (c) visualises the corresponding tree structure up to depth 4.

Exploratory modelling is a rapidly evolving branch of the model-based decision support literature. Starting from early conceptual ideas (Bankes, 1993; Lempert, 2002) over initial technical approaches (Groves and Lempert, 2007; Davis et al., 2007) towards sophisticated multi-dimensional analyses (Jafino and Kwakkel, 2021; Bonham et al., 2022; Steinmann et al., 2024; Hadjimichael et al., 2024a), a constant evolution and refinement is evident. This evolution was accompanied by multiple frameworks collating and structuring the current state of the art (e.g., Pianosi et al., 2016; Kwakkel and Haasnoot, 2019; Bonham et al., 2025), along with software implementations for key

analytical tools (Bryant, 2014; Hadka et al., 2015; Kwakkel, 2017; Hadjimichael et al., 2020a). By providing a structured workflow for evaluating scenario discovery algorithms, we contribute to this ongoing formalisation and establishment of exploratory modelling as a scientific approach for environmental and socio-environmental decision support.

The importance of visualisations for decision support in complex systems has been emphasised by multiple authors (Guivarch et al., 2022; Hakanen et al., 2023; Osika et al., 2023; Hadjimichael et al., 2024b). In the context of scenario discovery, visual inspection has proven to be a useful approach for understanding the trade-offs between different policies (Quinn et al., 2017; Kwakkel, 2019; Steinmann et al., 2024). We therefore make visual inspection a key element of our workflow, specifically designing test shapes which are amenable to visual inspection by virtue of their low dimensionality, but can be made more complex and challenging for rule induction algorithms by adding dimensions and fuzziness, and providing a suite of visualisations for inspecting and comparing scenario discovery algorithms.

A number of modifications to scenario discovery have been proposed (e.g., Lempert et al., 2008; Dalal et al., 2013; Kwakkel and Jaxa-Rozen, 2016; Hadjimichael et al., 2020b; Jafino and Kwakkel, 2021). However, a survey of the literature seems to show that few of these improvements have been taken up at large scale, with the notable exception of a new objective function for PRIM (Kwakkel and Jaxa-Rozen, 2016) by dint of becoming the default implementation in the most widely used software package. We hope that by providing a structured workflow and assets for evaluating such modifications, the technical underpinnings of scenario discovery can be improved in a more methodical manner.

A key element in our workflow is the three "classic" criteria for scenario discovery: coverage, density, and interpretability. For the first two, clear definitions and implementations are available. However, the third criterion, interpretability, has been highlighted by multiple authors as being somewhat ill-defined and thus requiring proxies (Kwakkel and Jaxa-Rozen, 2016; Kwakkel, 2019; Bonham et al., 2025). During the development of this workflow, we identified multiple proxies for interpretability, some of which are described in this paper. However, there is a need for future research on improved metrics for interpretability, preferably metrics which are agnostic to a specific choice of rule induction algorithm. This would enable a more principled comparison of different scenario discovery algorithms using unique interpretability proxies.

By establishing a clear process and language with which to talk about scenario discovery algorithms, we can also more readily start to explore the future evolution of scenario discovery. We observe a growing interest in multi-class scenario discovery work (Gerst et al., 2013; Rozenberg et al., 2014; Steinmann et al., 2020; Jafino and Kwakkel, 2021; Kahagalage et al., 2024; Bonham et al., 2025), which brings scenario discovery closer to conventional scenario-based planning and may therefore be more approachable for decision makers familiar with scenario methods (Wright et al., 2013; Bryant and Lempert, 2010; Steinmann et al., 2025). Furthermore, a number of authors (Quinn et al., 2018; Gold et al., 2019; Lamontagne et al., 2019; Trindade et al., 2019; Hadjimichael et al., 2020b) have investigated logistic regression as a tool for bringing scenario discovery closer to the intuitive understanding of cause-and-effect in complex systems, although the benefits and drawbacks compared to PRIM are currently unclear. Progress in either of these directions will be contingent on a robust foundation with which to evaluate potential iterations. Ultimately, analytical tools must clarify uncertainty, not add complexity. Our workflow supports this goal by providing a practical toolkit for the rigorous and transparent evaluation of future methods, strengthening the evidentiary basis for policy analysis.

The benefits of our structured approach are clearly visible in our case study on HHCART(D). Despite its conceptually more effective design, this algorithm did not outperform PRIM on our test shapes. Where performance was comparable, it came at the cost of more elaborate classification tree structures, which we proxy as low interpretability. The nominally oblique splits were, in practice, often still essentially axially aligned to the input parameter space, and the greedy implementation of the algorithm limits global coherence. It seems that the design objectives of classification algorithms (emphasising impurity reduction) are inherently at odds with the intended usage of scenario discovery (prioritising coherence and interpretability). Thus, it appears that PRIM is still the algorithm of choice for scenario discovery, mirroring findings in recent work on data mining algorithms (Arzamasov and Böhm, 2024). Any future tree-based algorithm which would improve on PRIM would need to emphasise global optimisation, node-level sparsity considerations, and structure-aware learning. It may also be that algorithms established in other domains such as data science (e.g., TURS by Yang and van Leeuwen, 2022), manufacturing (e.g., SIRUS by Bénard et al., 2021), or machine learning (as surveyed by Guidotti et al., 2018) are also worth evaluating in the context of scenario discovery, for which our proposed workflow provides the structure. This would also serve to validate (and improve) our procedure with algorithms using fundamentally different structures and metrics.

6. Conclusion

We presented a structured, configurable and extensible workflow for evaluating rule induction algorithms for scenario discovery. This workflow is accompanied by key metric implementations, customisable test shapes, and rich visualisation tools for inspecting analysis outcomes. We demonstrated this workflow using a comparison of PRIM, CART and HHCART(D), an oblique decision tree algorithm, and found that HHCART(D) does not yield substantial benefits over PRIM and CART for scenario discovery.

Our contributions establish a common ground for improving and evolving scenario discovery. Going forward, researchers may use our workflow, metrics, and test shapes to evaluate new algorithmic approaches for scenario discovery. This will make research outcomes more generalisable and complementary, strengthening our knowledge base on model-based decision support. Ultimately, this will allow us to continually evolve scenario discovery as a decision support tool for tackling complex socio-environmental challenges.

CRediT authorship contribution statement

Jasper T. ter Horst: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Conceptualization. **Patrick Steinmann:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization. **Jan H. Kwakkel:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization.

Software availability

The Python code and supporting datasets for this work are available on GitHub:

Workflow for Scenario Discovery Algorithms

The repository was created in 2025 by Jasper T. ter Horst. All materials are openly available under the *BSD 3-Clause New (Revised) License*. Please check the repository for the specific license terms before re-using the code or data.

Declaration on the use of generative AI and AI-assisted technologies

In preparing this work, we used ChatGPT (OpenAI) to assist with coding, writing, and reviewing. All AI-generated content was reviewed, edited, and integrated by the authors, who take full responsibility for the final content.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

All authors declare no conflicts of interest.

Acknowledgements

We thank Martijn Warnier for thoughtful feedback on the final draft of this work, which improved the clarity and presentation of this manuscript. We also thank the anonymous reviewers for constructive comments and suggestions that strengthened our work.

References

Arzamasov, V., Böhm, K., 2024. A reproducibility study of subgroup discovery algorithms. In: European Conference on Advances in Databases and Information Systems. Springer, pp. 3–13.

Auping, W.L., Pruyt, E., Kwakkel, J.H., 2015. Societal ageing in the Netherlands: A robust system dynamics approach. Syst. Res. Behav. Sci. 32 (4), 485–501.

Bankes, S., 1993. Exploratory modeling for policy analysis. Oper. Res. 41 (3), 435–449.

Bénard, C., Biau, G., Da Veiga, S., Scornet, E., 2021. SIRUS: Stable and interpretable RUle set for classification. Electron. J. Stat. 15, 427–505.

Birnbaum, A., Lamontagne, J., Wild, T., Dolan, F., Yarlagadda, B., 2022. Drivers of future physical water scarcity and its economic impacts in latin america and the caribbean. Earth's Futur. 10 (8), e2022EF002764.

Bonham, N., Kasprzyk, J., Zagona, E., 2022. Post-MORDM: Mapping policies to synthesize optimization and robustness results for decision-maker compromise. Environ. Model. Softw. 157, 105491.

Bonham, N., Kasprzyk, J., Zagona, E., 2025. Taxonomy of purposes, methods, and recommendations for vulnerability analysis. Environ. Model. Softw. 183, 106269.

Box, G.E., Hunter, J.S., Hunter, W.G., 2005. Statistics for Experimenters: Design, Innovation, and Discovery. John Wiley & Sons.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Routledge.

Bryant, B., 2014. Sdtoolkit: Scenario discovery tools to support robust decision making (v2. 33-1). Retrieved from cran.rproject.org/web/packages/sdtoolkit/index.html.

Bryant, B.P., Lempert, R.J., 2010. Thinking inside the box: A participatory, computer-assisted approach to scenario discovery. Technol. Forecast. Soc. Change 77 (1), 34–49.

Cañete-Sifuentes, L., Monroy, R., Medina-Pérez, M.A., 2021. A review and experimental comparison of multivariate decision trees. IEEE Access 9, 110451–110479.

Dalal, S., Han, B., Lempert, R., Jaycocks, A., Hackbarth, A., 2013. Improving scenario discovery using orthogonal rotations. Environ. Model. Softw. 48, 49–64.

Davis, P.K., Bankes, S.C., Egner, M., 2007. Enhancing Strategic Planning with Massive Scenario Generation: Theory and Experiments. Vol. 392, Rand Corporation.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Edali, M., 2022. Pattern-oriented analysis of system dynamics models via random forests. Syst. Dyn. Rev. 38 (2), 135–166.

Friedman, J.H., Fisher, N.I., 1999. Bump hunting in high-dimensional data. Stat. Comput. 9 (2), 123–143.

Führer, K., Kwakkel, J.H., d'Hont, F.M., Rouwette, E.A., Daalen, C.E.v., 2025. Towards participatory decision-making under deep uncertainty: Benefits and challenges. Int. J. Technol. Policy Manag. 25 (2), 150–173.

Gerst, M.D., Wang, P., Borsuk, M.E., 2013. Discovering plausible energy and economic futures under global change using multidimensional scenario discovery. Environ. Model. Softw. 44, 76–86.

Gold, D., Reed, P., Trindade, B., Characklis, G., 2019. Identifying actionable compromises: Navigating multi-city robustness conflicts to discover cooperative safe operating spaces for regional water supply portfolios. Water Resour. Res. 55 (11), 9024–9050.

Götz, P., Auping, W.L., Hinrichs-Krapels, S., 2024. Contributing to health system resilience during pandemics via purchasing and supply strategies: An exploratory system dynamics approach. BMC Health Serv. Res. 24 (1), 130.

Greeven, S., Kraan, O., Chappin, É.J., Kwakkel, J.H., 2016. The emergence of climate change mitigation action by society: An agent-based scenario discovery study. J. Artif. Soc. Soc. Simul. 19 (3).

Groves, D.G., Lempert, R.J., 2007. A new analytic method for finding policy-relevant scenarios. Glob. Environ. Chang. 17 (1), 73–85.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51 (5), 1–42.

Guivarch, C., Le Gallic, T., Bauer, N., Fragkos, P., Huppmann, D., Jaxa-Rozen, M., Keppo, I., Kriegler, E., Krisztin, T., Marangoni, G., et al., 2022. Using large ensembles of climate change mitigation scenarios for robust insights. Nat. Clim. Chang. 12 (5), 428–435.

Guivarch, C., Rozenberg, J., Schweizer, V., 2016. The diversity of socio-economic pathways and CO2 emissions scenarios: Insights from the investigation of a scenarios database. Environ. Model. Softw. 80, 336–353.

Haasnoot, M., Middelkoop, H., Van Beek, E., Van Deursen, W., 2011. A method to develop sustainable water management strategies for an uncertain future. Sustain. Dev. 19 (6), 369–381.

Hadjimichael, A., Gold, D., Hadka, D., Reed, P., 2020a. Rhodium: Python library for many-objective robust decision making and exploratory modeling. J. Open Res. Softw. 8.

Hadjimichael, A., Quinn, J., Wilson, E., Reed, P., Basdekas, L., Yates, D., Garrison, M., 2020b. Defining robustness, vulnerabilities, and consequential scenarios for diverse stakeholder interests in institutionally complex river basins. Earth's Futur. 8 (7), e2020EF001503.

Hadjimichael, A., Reed, P.M., Quinn, J.D., Vernon, C.R., Thurber, T., 2024a. Scenario storyline discovery for planning in multi-actor human-natural systems confronting change. Earth's Futur. 12 (9), e2023EF004252.

Hadjimichael, A., Schlumberger, J., Haasnoot, M., 2024b. Data visualisation for decision making under deep uncertainty: Current challenges and opportunities. Environ. Res. Lett. 19 (11), 111011.

Hadka, D., Herman, J., Reed, P., Keller, K., 2015. An open source framework for many-objective robust decision making. Environ. Model. Softw. 74, 114–129.

Hakanen, J., Gold, D., Miettinen, K., Reed, P.M., 2023. Visualisation for decision support in many-objective optimisation: State-of-the-art, guidance and future directions. In: Many-Criteria Optimization and Decision Analysis: State-of-the-Art, Present Challenges, and Future Perspectives. Springer, pp. 181–212.

Halim, R.A., Kwakkel, J.H., Tavasszy, L.A., 2016. A scenario discovery study of the impact of uncertainties in the global container transport system on European ports. Futures 81, 148–160.

Jafino, B.A., Kwakkel, J.H., 2021. A novel concurrent approach for multiclass scenario discovery using multivariate regression trees: Exploring spatial inequality patterns in the Vietnam Mekong Delta under uncertainty. Environ. Model. Softw. 145, 105177.

Kahagalage, S.D., Turan, H.H., Elsawah, S., Gary, M.S., 2024. Exploratory modelling and analysis to support decision-making under deep uncertainty: A case study from defence resource planning and asset management. Technol. Forecast. Soc. Change 200, 123150.

Kozlova, M., Moss, R.J., Yeomans, J.S., Caers, J., 2024. Uncovering heterogeneous effects in computational models for sustainable decision-making. Environ. Model. Softw. 171, 105898.

Kwakkel, J.H., 2017. The exploratory modeling workbench: An open source toolkit for exploratory modeling, scenario discovery, and (multi-objective) robust decision making. Environ. Model. Softw. 96, 239–250.

Kwakkel, J.H., 2019. A generalized many-objective optimization approach for scenario discovery. Futures Foresight Sci. 1 (2), e8.

Kwakkel, J.H., Auping, W.L., Pruyt, E., 2013. Dynamic scenario discovery under deep uncertainty: The future of copper. Technol. Forecast. Soc. Change 80 (4), 789–800.

Kwakkel, J.H., Cunningham, S.C., 2016. Improving scenario discovery by bagging random boxes. Technol. Forecast. Soc. Change 111, 124–134.

Kwakkel, J.H., Haasnoot, M., 2019. Supporting DMDU: A taxonomy of approaches and tools. In: Marchau, V.A.W.J., Walker, W.E., Bloemen, P.J.T.M., Popper, S.W. (Eds.), Decision Making under Deep Uncertainty. Springer International Publishing, Cham, pp. 355–374.

Kwakkel, J.H., Jaxa-Rozen, M., 2016. Improving scenario discovery for handling heterogeneous uncertainties and multinomial classified outcomes. Environ. Model. Softw. 79, 311–321.

Lamontagne, J., Reed, P., Marangoni, G., Keller, K., Garner, G., 2019. Robust abatement pathways to tolerable climate futures require immediate global action. Nat. Clim. Chang. 9 (4), 290–294.

Lempert, R.J., 2002. A new decision sciences for complex systems. Proc. Natl. Acad. Sci. 99 (suppl_3), 7309–7313.

Lempert, R.J., 2003. Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis. Technical Report MR-1626, RAND Corporation.

Lempert, R.J., Bryant, B.P., Bankes, S.C., 2008. Comparing Algorithms for Scenario Discovery. Technical Report WR-557-NSF, RAND Corporation.

Lempert, R.J., Groves, D.G., Popper, S.W., Bankes, S.C., 2006. A general, analytic method for generating robust strategies and narrative scenarios. Manag. Sci. 52 (4), 514–528.

Lipton, Z.C., 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16 (3), 31–57.

Majumder, T., 2020. Ensembles of Oblique Decision Trees Master's thesis. The University of Texas at Dallas, Richardson, TX.

Manheim, D., 2023. Building less-flawed metrics: Understanding and creating better measurement and incentive systems. Patterns 4 (10).

Merino-Benítez, T., Bojórquez-Tapia, L.A., Miquelajauregui, Y., Batllori-Sampedro, E., 2024. Navigating climate change complexity and deep uncertainty: Approach for building socio-ecological resilience using qualitative dynamic simulation. Front. Clim. 6, 1331945.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592.

Murthy, S.K., Kasif, S., Salzberg, S., 1994. A system for induction of oblique decision trees. J. Artificial Intelligence Res. 2, 1–32.

Oostdijk, M., Elsler, L.G., Van Deelen, J., Auping, W.L., Kwakkel, J., Schadeberg, A., Vastenhoud, B.M., Nedelciu, C.E., Berzaghi, F., Prellezo, R., et al., 2024. Modeling fisheries and carbon sequestration ecosystem services under deep uncertainty in the ocean twilight zone. Ambio 53 (11), 1632–1648.

Osika, Z., Salazar, J.Z., Roijers, D.M., Oliehoek, F.A., Murukannaiah, P.K., 2023. What lies beyond the Pareto front? A survey on decision-support methods for multi-objective optimization. arXiv preprint arXiv:2311.11288.

Parker, A.M., Srinivasan, S.V., Lempert, R.J., Berry, S.H., 2015. Evaluating simulation-derived scenarios for effective decision support. Technol. Forecast. Soc. Change 91, 64–77.

Pianosi, F., Beven, K., Freer, J., Hall, J.W., Rougier, J., Stephenson, D.B., Wagener, T., 2016. Sensitivity analysis of environmental models: A systematic review with practical workflow. Environ. Model. Softw. 79, 214–232.

Quinn, J.D., Reed, P.M., Giuliani, M., Castelletti, A., Oyler, J.W., Nicholas, R.E., 2018. Exploring how changing monsoonal dynamics and human pressures challenge multireservoir management for flood protection, hydropower production, and agricultural water supply. Water Resour. Res. 54 (7), 4638–4662.

Quinn, J.D., Reed, P.M., Keller, K., 2017. Direct policy search for robust multi-objective management of deeply uncertain socio-ecological tipping points. Environ. Model. Softw. 92, 125–141.

Rozenberg, J., Guivarch, C., Lempert, R., Hallegatte, S., 2014. Building SSPs for climate policy analysis: A scenario elicitation methodology to map the space of possible future challenges to mitigation and adaptation. Clim. Change 122 (3), 509–522.

Sassi, O., Crassous, R., Hourcade, J.-C., Gitz, V., Waisman, H., Guivarch, C., 2010. IMACLIM-R: A modelling framework to simulate sustainable development pathways. Int. J. Glob. Environ. Issues 10 (1–2), 5–24.

Schlumberger, J., Gold, D., Di Fant, V., Winter, G., Taner, M.Ü., Kwakkel, J., 2026. A review of tools and resources to support decision-making under deep uncertainty. Environ. Model. Softw. 106900.

Schwartz, P., 1997. Art of the Long View: Planning for the Future in an Uncertain World. John Wiley & Sons.

Steinmann, P., Auping, W.L., Kwakkel, J.H., 2020. Behavior-based scenario discovery using time series clustering. Technol. Forecast. Soc. Change 156, 120052.

Steinmann, P., Verstegen, J., Van Voorn, G., Roman, S., Ligtenberg, A., 2025. Scenario search: Finding diverse, plausible and comprehensive scenario sets for complex systems. Socio-Environ. Syst. Model. 7, 18823.

Steinmann, P., van der Zwet, K., Keijser, B., 2024. Simulation-based generation and analysis of multidimensional future scenarios with time series clustering. Futur. Foresight Sci. 6 (4), e194.

Student, J., Kramer, M.R., Steinmann, P., 2020. Simulating emerging coastal tourism vulnerabilities: An agent-based modelling approach. Ann. Tour. Res. 85, 103034.

Sücüllü, C., Yücel, G., 2014. Behavior analysis and testing software (BATS). In: Proceedings of the 32nd International Conference of the System Dynamics Society. International System Dynamics Society, pp. 20–24.

Sunkara, S.V., Singh, R., Gold, D., Reed, P., Bhave, A., 2023. How should diverse stakeholder interests shape evaluations of complex water resources systems robustness when confronting deeply uncertain changes? Earth's Futur. 11 (8), e2022EF003469.

ter Horst, J.T., 2025. Thinking Outside the Box: A Critical Evaluation of Oblique Decision Tree Algorithms for Scenario Discovery Master's thesis. Delft University of Technology, Faculty of Technology, Policy and Management, Engineering and Policy Analysis.

Toman, M.A., Lempert, R.J., 2008. Impacts on US Energy Expenditures and Greenhouse-Gas Emissions of Increasing Renewable-Energy Use: Technical Report. Vol. 384, Rand Corporation.

Trindade, B., Reed, P., Characklis, G., 2019. Deeply uncertain pathways: Integrated multi-city regional water supply infrastructure investment and portfolio management. Adv. Water Resour. 134, 103442.

van Droffelaar, I.S., 2020. A Dependent Sampling Approach to Scenario Discovery Master's thesis. Delft University of Technology, Faculty of Technology, Policy and Management, Engineering and Policy Analysis.

Vrugt, J.A., 2016. Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. Environ. Model. Softw. 75, 273–316.

Wickramarachchi, D.C., Robertson, B.L., Reale, M., Price, C.J., Brown, J., 2016. HHCART: An oblique decision tree. Comput. Statist. Data Anal. 96, 12–23.

Wright, G., Bradfield, R., Cairns, G., 2013. Does the intuitive logics method – and its recent enhancements – produce "effective" scenarios? Technol. Forecast. Soc. Change 80 (4), 631–642.

Yang, L., van Leeuwen, M., 2022. Truly unordered probabilistic rule sets for multi-class classification. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 87–103.