A simulation-based approach for reconstructing a diverse set of supply chain models with sparse data using a quality diversity algorithm

van Schilt, Isabelle M.; Kwakkel, Jan H.; Mense, Jelte P.; Verbraeck, Alexander

DOI
10.1016/j.simpat.2025.103216

Publication date
2026

Document Version
Final published version

Published in
Simulation Modelling Practice and Theory

Simulation Modelling Practice and Theory 146 (2026) 103216

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

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

A simulation-based approach for reconstructing a diverse set of supply chain models with sparse data using a quality diversity algorithm

Isabelle M. van Schilt [a], Jan H. Kwakkel [a], Jelte P. Mense [b], Alexander Verbraeck [a],*

[a] Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, Delft, 2628BX, Netherlands
[b] Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, Delft, 2628CD, Netherlands

ARTICLE INFO

Keywords:
Simulation
Quality-diversity
Data sparseness
Supply chain

ABSTRACT

Data on supply chains is often sparse due to reluctance among actors to share their data, making supply chain simulation modeling difficult. As a result, supply chain simulation models suffer from parametric and structural uncertainties, and there is a large variety of plausible simulation models that would align with the sparse observations about the real-world supply chain. Constructing a diverse set of models that fit sparse data is not an easy task. A relatively unknown approach to generating this diverse set of plausible models is the Quality Diversity (QD) algorithm. This study evaluates the feasibility of using QD to generate a diverse ensemble of supply chain simulation models for a varying degree of data sparseness. The results show that QD is able to generate a diverse ensemble of supply chain models, including the ground truth. As expected, QD successfully identifies the structure of the ground truth most frequently for a low level of data sparseness. When the sparseness of the data increases, QD is prone to overfitting, identifying supply chain structures that are more complex than the ground truth. Further research should focus on reviewing the calibration metric for sparse data, to reduce the overfitting of complex network structures.

1. Introduction

During COVID-19, there was a steep rise in demand for Personal Protective Equipment (PPE), such as goggles, gloves, face masks, and respirators [1]. Many countries suddenly needed a large supply of medical PPE to protect caretakers in hospitals. This high demand, unfortunately, also led to an opportunity for the illicit market to produce and distribute counterfeit PPE [2]. Fraudulent organizations were, for instance, selling non-medical PPE as medical PPE with a high profit margin, leading to unacceptable health risks [3,4]. Due to the lack of historical data on COVID-19 and criminals trying to obfuscate their data as much as possible, the illicit activities and related logistics of these fraudulent organizations were mostly invisible [5]. This made it difficult for law enforcement to intervene effectively, seize counterfeit PPE, and stop crime.

Simulation is a way to understand a system and to measure the effectiveness changes to the system by modeling the system's behavior over time [6,7]. This paper focuses on discrete event simulation models to represent illicit supply chains [8,9]. A discrete event simulation model consists of components, their parameters, and their behavior over time, as well as the relations between the components. For example, a simulation model of a supply chain consists of actors with parameters such as the time to process a

* Corresponding author.
E-mail address: a.verbraeck@tudelft.nl (A. Verbraeck).

https://doi.org/10.1016/j.simpat.2025.103216
Received 19 July 2024; Received in revised form 22 September 2025; Accepted 19 October 2025

Available online 28 October 2025

product, inventory levels, and transportation times. The relations define the connections between the actors to represent the network, i.e., the structure, of the supply chain. Calibration uses data to tune the parameters of a simulation model in such a way that the model behavior sufficiently matches the system behavior in the real world [10–12].

Data on supply chains, such as demand, inventory levels, processing times, or transportation times, is often sparse due to reluctance among supply chain actors to share their (correct) data [13,14]. This reluctance has various causes, such as competition, high data cost, or illicit activities in the case of fraudulent supply chains [15]. When data is sparse, the probability of equifinality during the calibration process increases. Multiple versions of the supply chain simulation model can correctly reproduce the sparsely observed data [5]. Therefore, a system with sparse data cannot be fully explained by a single theory or model but needs a variety of models [16,17]. This is in line with the many-model thinking approach of Page [18] that emphasizes the need for an ensemble of models to understand and analyze a complex system.

However, creating a useful ensemble of models based on sparse data is not easy. Many calibration algorithms that can generate multiple model configurations converge to similar solutions, as configurations with parameters that differ slightly from the best-matching solution typically outperform those with vastly different parameters [18]. Finding a *diverse* set of, say, 20 solutions during calibration is therefore harder than finding the top-20 best solutions, yet a diverse set is more desirable [19].

One of the approaches for generating a diverse set of plausible models is using a Quality Diversity (QD) algorithm. QD algorithms use evolutionary concepts to find optimal solutions at multiple points of the user-defined search space [20,21]. QD is mostly used in the field of robotics and reinforcement learning [22–24]. Recent work applied QD for multi-objective optimization [25], hyperparameter tuning of a machine learning model [26], and for identifying the most preferred solutions to decision-makers [27]. However, QD remains unexplored in many other application areas, since it is a relatively new approach for evolutionary computation [22,26]. Our main research question is whether QD algorithms can generate a diverse set of plausible configurations of supply chain simulation models characterized by sparse data.

First, we review related work on generating a diverse set of simulation models that can be calibrated with sparse data, and on QD algorithms. Next, we assess the feasibility of a QD algorithm for generating a diverse set of supply chain configurations that can replicate the observed (sparse) data. To test the approach, we use a ground truth simulation model of a synthetic counterfeit PPE supply chain. For the analysis, we extract data from the ground truth model and vary the degree of data sparseness. Next, we assess whether the QD algorithm can generate the ground truth as a feasible solution among its diverse set of solutions. Hence, our study offers a first insight into the potential of using QD algorithms to generate an ensemble of diverse and plausible configurations of simulation models, particularly for supply chains with sparse data. Such an ensemble of plausible configurations, in turn, enables decision-makers to make more robust decisions on, for example, system changes that are effective for the *ensemble* of supply chain models, rather than for a single simulation model based on the observed data.

The paper is structured as follows. Section 2 presents the relevant state-of-the-art literature. Section 3 describes the method for evaluating the application of the QD algorithm, as well as the used case study. Section 4 shows the results of using the QD algorithm in a simulation context. Section 5 discusses our results. Section 6 concludes our work and presents further research.

2. Literature review

This section presents the current state-of-the-art for our research. First, we show the related work on building and calibrating simulation models using sparse data. Second, we position our paper in relation to the literature on a pluralist view on simulation modeling. Third, we examine the related work on QD and provide more insight into the QD algorithm itself.

2.1. Simulation with sparse data

Simulating a real-world system becomes challenging when data is sparse [28]. Data sparseness makes it more difficult to accurately mimic the behavior of the real world in a simulation model [5,29,30]. Several studies have examined the calibration of simulation models to real-world sparse data.

Among the first authors to explicitly address the calibration of a simulation model when data is sparse is Liu et al. [31]. They propose a simulation–optimization approach to automatically tune the parameters of an agent-based simulation model using sparse data for an emergency department. The problem is formulated as a series of local minimum search problems. Vanbrabant et al. [29] present a framework for assessing real-world input data quality problems for emergency department simulation models. De Santis et al. [32] focus on the calibration of a discrete event simulation model under data sparseness. Observable values from the real-world system are used to determine the parameter values of the simulation model, for example, for time intervals. de Groot and Hübl [33] use calibration as a validation method; in their case, the sparseness of data makes validating the simulation model challenging. They manually fine-tune the parameters and dynamics of the model to enhance validity. Srikrishnan and Keller [28] calibrate an agent-based simulation model on housing abandonment under flood risk, and show that limited data can be insufficient for correctly identifying the model structure. van Schilt et al. [5] compare various calibration techniques for simulation models when increasing the degree of data sparseness. They calibrate the parameters of a discrete event simulation model of a counterfeit PPE supply chain. Using the same case study, van Schilt et al. [14] evaluate the effect of three dimensions of data sparseness (noise, bias, and missing values) on supply chain visibility using simulation. van Schilt et al. [34] compares four calibration methods (Powell's Method, Approximate Bayesian Computing, Genetic Algorithm, and Bayesian Optimization) for identifying the structure of the supply chain given an increasing degree of data sparseness.

2.2. A pluralist view on simulation modeling

When modeling a complex phenomenon characterized by sparse data and uncertainty, equifinality among the simulation models may exist. This means that multiple plausible simulation models are coherent with the available sparse data of the real world. Only focusing on one model to analyze the system could lead to a "wrong" view on the phenomenon, and hence render the model ineffective [5]. Often, a complex phenomenon, such as a supply chain, cannot be captured by a single theory or model when data is sparse [18]. Illicit supply chains are a good example of supply chains where data is intentionally sparse, but regular supply chains also suffer from incomplete and erroneous data. Identifying the structure and parameter values of a supply chain for simulation purposes using sparse data is typically a case where a pluralist view on simulation modeling can be useful.

Pluralism as a research philosophy refers to a diversity of views, theories, or models that are required to explain a complex phenomenon, rather than using just a single view, theory, or model. Weisberg [35] argues that the more models are available, the more likely robust properties amongst the models can be found that can be related to the real world. Similarly, Durán and Formanek [19] note that a heterogeneous ensemble of models is needed for robustness analysis. Veit [17] takes these statements even further and argues that (1) any successful analysis should be focused on a target set of models, and (2) for almost any aspect of a phenomenon, scientists require multiple models to achieve a goal. As every single simulation model is a limited representation of the world, it only gives one view on the system [36]. This is especially undesirable for systems analysis when there is so little data available about a model property, that it is not even possible to use a probability distribution, a phenomenon also known as deep uncertainty [37]. From a modeling perspective, Thompson and Smith [38] state that having an ensemble of models reduces the possibility of errors. Simulation models require user-defined initial conditions based on real-world data. The idea is to generate an ensemble of models with different initial conditions to account for these errors in the initial conditions. Similarly, Page [18] notes that scientists often cannot use one single simulation model, and introduces the many-model approach. This approach refers to the need for multiple models to understand a complex system.

2.3. Quality diversity algorithms

Quality Diversity (QD) algorithms, or illumination algorithms, aim to find the *most diverse set* of close-to-optimal solutions using evolutionary concepts [20]. Traditional optimization algorithms aim to find the best solution within a specific search space, while illumination algorithms focus on providing the highest-performing solution at every user-defined point within that search space [20,21].

Table 1 shows a number of methods in addition to QD to reconstruct hidden states or parameters of models, namely calibration, data assimilation, inverse modeling, and Bayesian inference. Most methods predominantly focus on parametric uncertainty only, and emphasize either a single best-fitting solution or a probabilistic distribution of plausible solutions. While these methods are powerful, they often converge to one dominant explanation. In contrast, Quality Diversity algorithms keep a wide range of possible reconstructions. This algorithmic diversity does not provide probabilities, but instead shows a broad set of alternative explanations that remain consistent with the observations. Table 1 also highlights that QD is explicitly aimed at providing a *diverse* set of solutions.

2.3.1. Use of QD in modeling

QD algorithms are widely applied in the field of reinforcement learning and robotics [23,24]. A classic example is maze navigation where QD is used to generate a set of diverse behaviors for the robot movement to solve the maze [22,46–48]. The diverse behaviors are specified in the form of input parameters to the robot. No trivial mapping of the input parameters to the robot movement exists, making quality diversity interesting for these types of problems.

QD algorithms can be used to present a diverse set of solutions to decision-makers. Recent work of Kent and Branke [27] combines a quality diversity search with Bayesian optimization. Their approach efficiently identifies the most preferred solutions by including the decision-makers in the process.

The application of QD algorithms for model calibration is a relatively unexplored area. Schneider et al. [26] compare various quality diversity algorithms for hyperparameter optimization of a machine learning model. To our knowledge, no research has been performed on calibrating simulation models using QD.

2.3.2. Parameter spaces

QD algorithms are based on evolutionary concepts of "survival-of-the-fittest" [21,22]. The algorithms use three parameter spaces: (1) input space, (2) behavior space, and (3) output space. The input space includes the so-called genotypes x, also known as the input parameters. The behavior space includes the so-called phenotypes $b(x)$, also known as the dimensions describing the behavior of the input parameters x. The output space is the solution space and includes the model outputs, $f(x)$. Input space, behavior space, and model outputs that constitute the output space are shown in Fig. 1.

The general idea is that the QD algorithms create diversity by discretizing the behavior space, and fill each container in this space with an optimal solution for that container [21]. To do this, a mapping between the input space and the behavior space is needed. In a sense, a dimension reduction happens when mapping the genotypes x to the phenotypes $b(x)$. Direct encoding is a straightforward mapping from x to $b(x)$, which occurs when genotypes affect an independent parameter of the phenotypes. Indirect encoding occurs when a genotype affects multiple parameters of the phenotypes [20]. Choosing the mapping highly impacts the quality of the solutions, but unfortunately, it is quite challenging to define a good quality behavior space [45]. One way to define

Table 1
Comparison of reconstruction approaches for aligning simulations with observational data.

Method	Properties
Calibration	Kennedy and O'Hagan [39]
Objective:	Minimize the difference between simulation and observations
Output:	Single best-fitting parameter set
Solutions Diversity:	Low
Structural Uncertainty:	No, fixed model structure
Uncertainty Treatment:	Limited, often via sensitivity analysis
Data Assimilation	Evensen [40,41]
Objective:	Sequentially align model states with observational data
Output:	Updated trajectory or state estimate
Solutions Diversity:	Low–Moderate (ensemble spread)
Structural Uncertainty:	No, model structure assumed correct
Uncertainty Treatment:	Probabilistic, typically Gaussian
Inverse Modeling	Tarantola [42]
Objective:	Infer hidden inputs or forcing that reproduce observations
Output:	Optimized inputs or forcing
Solutions Diversity:	Low
Structural Uncertainty:	No, structure fixed; separate runs for alternative models
Uncertainty Treatment:	Usually deterministic; uncertainty explored post hoc
Bayesian Inference	Beaumont et al. [43] and Beaumont [44]
Objective:	Characterize parameter values consistent with data
Output:	Posterior distribution samples
Solutions Diversity:	Moderate (spread of posterior)
Structural Uncertainty:	Yes, via Bayesian model comparison, but hard to scale
Uncertainty Treatment:	Explicit probabilistic treatment
Quality Diversity	Mouret and Clune [20],Cully and Demiris [45]
Objective:	Explore diverse high-performing reconstructions
Output:	Archive of diverse solutions across behavioral dimensions
Solutions Diversity:	High (explicit diversity objective)
Structural Uncertainty:	Yes, exploration of parametric and structural variants in archive
Uncertainty Treatment:	Algorithmic diversity: systematic discovery, no probabilities

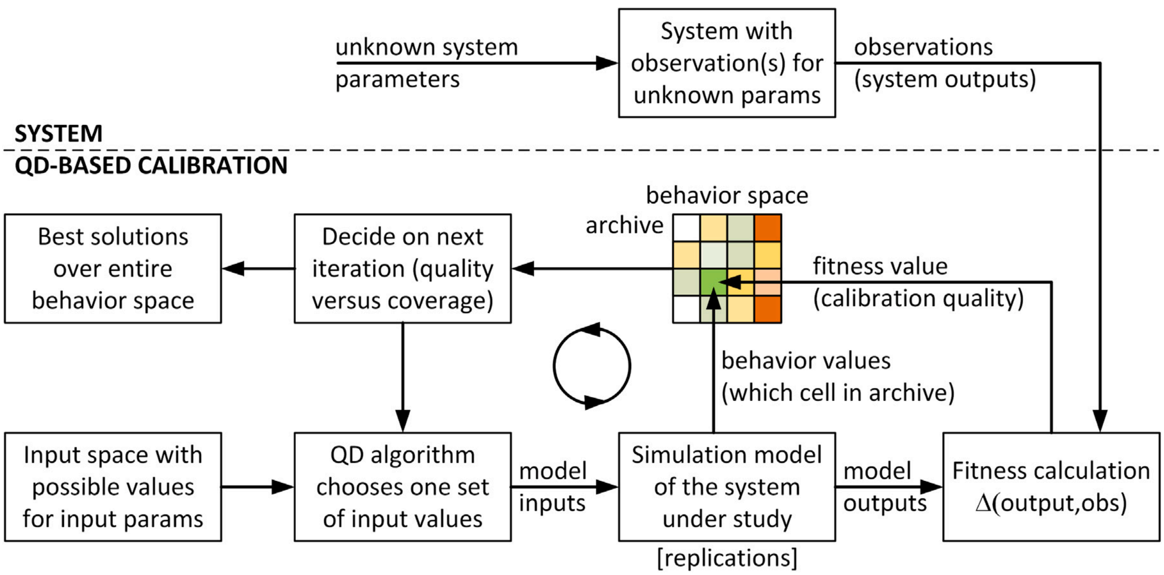


Fig. 1. Simplified diagram of the QD process.

the dimensions of the behavior space is to use techniques such as Principal Component Analysis. However, in many cases, defining the dimensions of the behavior space cannot be carried out automatically, and expert knowledge is required [21].

The behavior space needs to be discretized to identify subspaces or "containers" in which optimal solutions will be found. Discretization of the behavior space is often done by using a grid, see Fig. 1 for a simple example. A grid-based approach is easy to understand and implement for QD algorithms. A user-defined number of discrete intervals per dimension of the behavior space is needed to create the multi-dimensional hypercuboid containers using the grid. For high-dimensional behavior spaces, Vassiliades

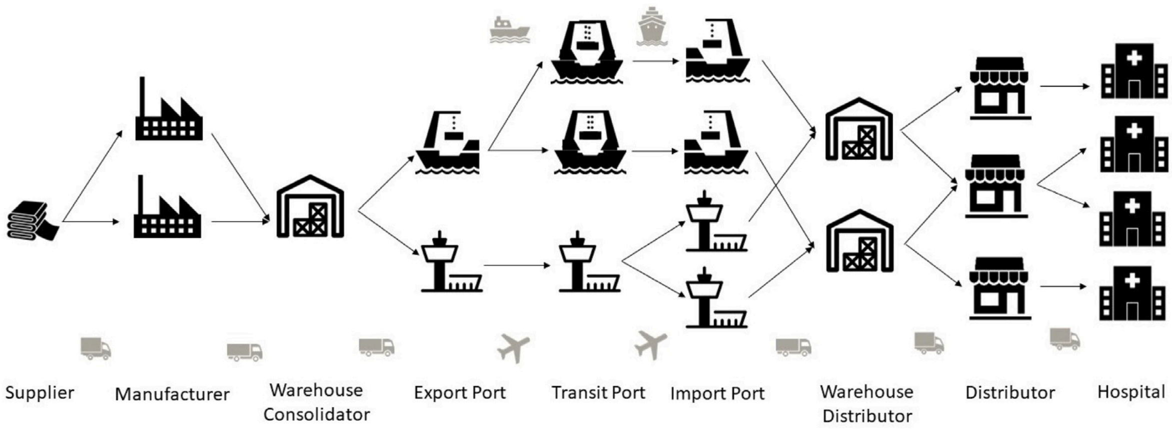


Fig. 2. Synthetic ground truth supply chain of counterfeit PPE.

et al. [49] introduces Centroidal Voronoi Tessellation (CVT) as a tool. CVT explicitly controls the number of containers, and the resulting containers have a convex polygonal shape with corresponding centroids.

2.3.3. QD algorithms

Finding the optimal solution for each container in the behavior space is done with a QD algorithm. One of the first and most widely applied QD algorithms is *MAP-Elites* [20]. MAP-Elites starts with generating a set of candidate solutions with randomly chosen genotypes x following a Gaussian distribution. The behavior $b(x)$ and the output $f(x)$ of these candidate solutions are calculated. Next, the candidate solutions are placed into the containers to which they belong in the behavior space. When multiple candidate solutions are placed in the same container, the highest-performing one (i.e., the best output) is kept. After initialization, the search algorithm starts. It randomly selects a container in the discretized behavior space. Mutation and crossover are used to generate offspring from the candidate solution in the container. If the offspring has a higher output value, it replaces the current candidate solution. This process repeats until a stopping criterion, e.g., the number of function evaluations, is reached.

Another commonly used QD algorithm is *Novelty Search with Local Competition*, introduced by Lehman and Stanley [50]. This algorithm compares the quality and the diversity of a candidate solution relative to its neighbor. It optimizes the candidate solutions on (1) quality: maximizing the output relative to its neighbors, and (2) diversity: maximizing the novelty objective on how far the solution in the behavior space is distant from its neighbors. The main limitation of this algorithm is that this algorithm creates two individual sets of solutions for quality and diversity. This is less effective than having the single set of solutions generated by MAP-Elites [21].

More recently, Fontaine et al. [51] proposed *Covariance Matrix Adaptation MAP-Elites (CMA-ME)*. This algorithm combines the popular MAP-Elites algorithm with a single-objective optimization algorithm called Covariance Matrix Adaptation Evolution Strategy (CMA-ES). This hybrid algorithm efficiently explores new areas in the search space using MAP-Elites, while using the selection and adaptation rules of CMA-ES to find high-quality solutions. Fontaine and Nikolaidis [48] shows that CMA-ME outperforms MAP-Elites for finding a diverse set of optimal solutions.

Another recent extension of MAP-Elites is *multi-objective MAP-Elites* of Pierrot et al. [25]. It builds on MAP-Elites, but uses multi-objective optimization, providing insights into trade-offs for this diverse set of Pareto optimal solutions.

The main limitation of QD algorithms is that they cannot guarantee that all the containers in the discretized behavior space are filled [20,21,50]. Since the discretization of the behavior space is a user-defined process, it can occur that the genotypes do not map to some phenotypes. Another limitation is that many candidate solutions with different genotypes could be present in the same container [20].

3. Method

This section outlines the method for generating and evaluating a diverse set of optimal supply chain configurations using a QD algorithm. First, the case study is discussed. Second, the QD algorithm is presented. Finally, the experiments that use a ground truth set-up are explained.

3.1. Case study

Based on open-source data and expert interviews, we use one specific configuration of a stylized counterfeit PPE supply chain as ground truth. The case study is fictitious and bears no direct relation to actual organizations or locations of counterfeit PPE production.

Fig. 2 visualizes the structure of the ground truth counterfeit PPE supply chain simulation model from China to the northeast United States of America (USA). The symbols in the figure represent the main actors in the supply chain, and the arrows represent the transportation flows. A description of the stylized supply chain is given in Appendix A.

We develop a discrete event simulation model of this stylized configuration of a counterfeit PPE supply chain. In the simulation model, most uncertainties such as processing times of actors, delays during transport, and speed of different modes of transport follow triangular distributions inspired by real-world data of a fashion retailer and expert interviews [2,52]. Table A.2 in the Appendix shows the input parameters for the actors and the links used in the ground truth simulation model. Table A.2 in the Appendix shows parametrization of the speed and the delays of the transport modalities for the simulation model of this study. The simulation model has been developed with the library *pydsol-core* and *pydsol-model* in Python in combination with *networkx*. The library *pydsol* is a Python implementation of the Distributed Simulation Object Library (DSOL), originally implemented in Java [53].

The ground truth discrete event simulation model is developed to produce the observed data of the system. We extract time series data from the simulation model that describes when a quantity of PPE arrives at an actor, including the location and the type of actor. For example, a batch with a quantity of 20.000 PPE arrives at the export airport in Hong Kong on day 3. Data of the time series is summed per day, and is aggregated over the actor types. Since the model is stochastic, it uses multiple replications; we run the model for a simulation time of 52 weeks with 10 replications with unique seeds.

3.2. Configuration of quality diversity algorithm

In this research, the QD algorithm Covariance Matrix Adaptation MAP-Elites (CMA-ME) of Fontaine et al. [51] is used for finding a diverse set of optimal solutions due to its high performance. Emitters are instances of the CMA-ME algorithm that generate new candidate solutions, adapt, and save the population of solutions. The algorithm is initialized with an emitter across ten unique seeds. Each emitter generates 96 candidate solutions in each iteration. A convergence analysis is performed on the number of quality improvements and the solution diversity, to determine the number of iterations required. The quality diversity algorithm is implemented using the python library *pyribs* [24].

The configuration of the QD algorithm asks for the description of the following three spaces: the input space, the behavior space, and the output space.

Input space. The input space of the QD algorithm defines the parameters to calibrate. These are uncertain parameters in the simulation model that need to be tuned such that the model's behavior matches the real-world behavior. In the case of illicit supply chains, both the *structure* and the *parameters* of the supply chain simulation model are uncertain. Table A.3 in the Appendix gives an overview of the defined input space features and the corresponding input space parameters. Each parameter in the input space has bounds to ensure that the algorithm chooses feasible candidate solutions. For example, a simulation model cannot schedule an event in the past, so a parameter related to time cannot be lower than zero. Note the parameters excluded from a composition's input space are considered known and correspond to the ground truth.

To give a clear definition of the terminology, the simulation model itself has *input parameters* that are either drawn from a (known) distribution, or fixed. The QD input space has a number of *input space features* that correspond to the simulation model's input parameters, where an input space feature can have multiple *input space parameters*. For the analysis of the case study, the input space features are grouped into six *feature profiles*: structure, modus operandi, source, legal, import, and destination. The input parameters for the simulation model can be found in the Appendix in Tables A.1 and A.2, and the input space profiles, features, parameters and bounds can be found in Table A.3. The input space parameters form the dimensions of the high-dimensional input space.

The first feature of the input space is the structure of the supply chain. For the structural uncertainty, a set of 40,000 possible supply chain configuration graphs with different numbers of actors and connectivity have been generated using the System Entity Structure approach [54,55]. The set of graphs is ranked on the density in each graph, and the graph index is part of the input space. As explained above, bounds are set for the input space parameters of all other features as well.

Behavior space. The behavior space defines the dimensions on which the diverse solutions are positioned. The input space is mapped to the behavior space to ensure that the chosen sets of input parameters for the simulation model describe a wide variety of behavior. For illicit supply chains and legal supply chains alike, it is interesting to find diverse and optimal supply chain simulation models with different transport costs and various degrees of network vulnerability [30]. Transport costs are an essential part of profit-driven crime; lower costs mean more profit [56]. Network resilience shows the extent to which the network is vulnerable to interventions of law enforcement [15]. Both these dimensions are important for legal supply chains as well, since most organizations strive for a supply chain with low transport costs and resilience against disruptions.

As the next step in the QD method, the 27 input parameters in the input space have to be mapped to the two chosen dimensions in the behavior space. The transport cost in the behavior space is defined as the average total transport cost of a product, and is dependent on many of the input parameters such as choice of transport mode, transport distances, and used ports or airports. For the network resilience dimension, Gao et al. [57] state that density, i.e., the ratio of the actual number of edges to the maximum possible number of edges in a network, is one of the key factors influencing a network's resilience. Therefore, we use network density as a measure of network resilience in the behavior space. This maps the graph structure from the input space on the resilience dimension of the behavior space. The exact calculations are described in Appendix A.3.

The behavior space is discretized into a grid of 10 x 10 containers. We refer to these as QD containers. For the case study, the range of transport costs in this behavior space is between $250 and $1250, and, for density, between 0.02 and 0.07. The behavior space has only two dimensions, enhancing the understandability and interpretability of the results [21].

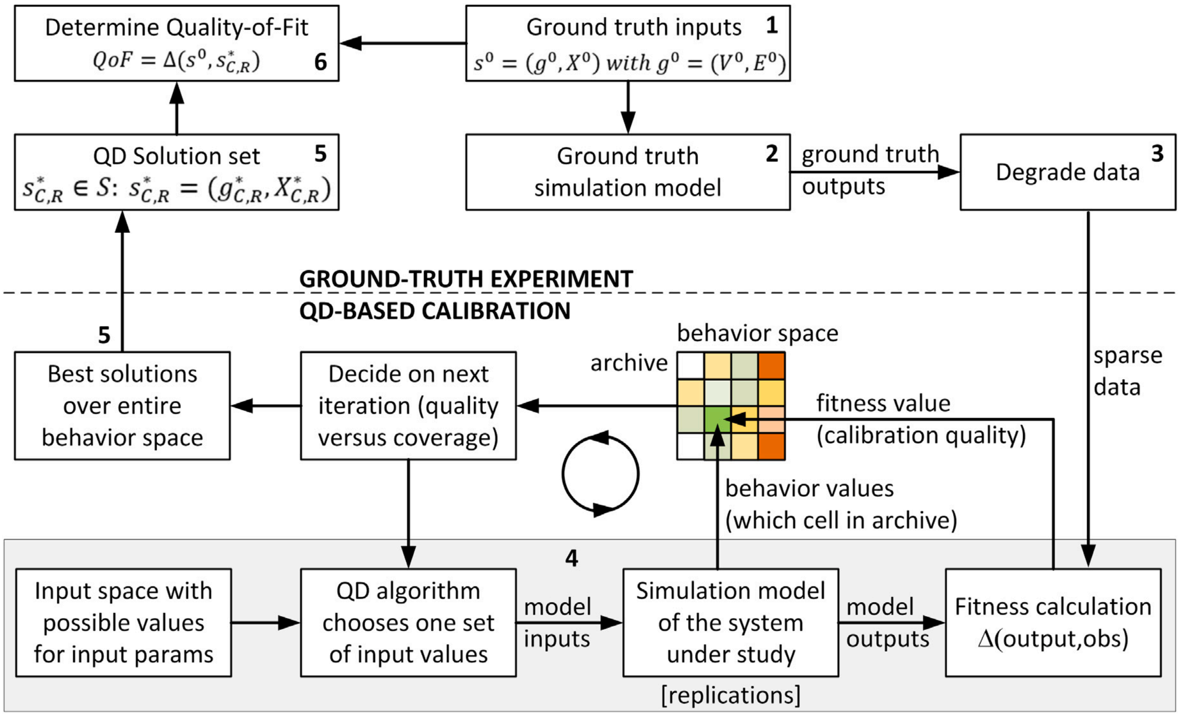


Fig. 3. Ground truth experiment set-up for QD.

Output space. The output space defines the objective to minimize, in this case, *the distance between the ground truth output and the output of the candidate solution*. A distance metric is used to describe the distance between the simulation model data and the observed data given a certain function. In this research, we use a classic distance metric: the Manhattan (L1) distance. The Manhattan distance is the sum of the absolute differences for each dimension of the data points. This distance metric is highly efficient for complex and high-dimensional data applications such as discrete event simulation models [58,59]. In our research, we compare the aggregated time series data of each actor resulting from the ground truth simulation model and the candidate simulation model. We normalize the Manhattan distance of each actor using the 5th percentile and 95th percentile of the actor's ground truth data. Next, we sum the normalized Manhattan distance of each actor to get the overall Manhattan distance between the ground truth and candidate solution.

3.3. Design of experiments

For our experiments, a ground truth set-up is used to assess the feasibility of applying QD under a varying degree of data sparseness. This set-up allows us to measure how closely QD calibrates the "true" values, which is challenging when dealing with real-world data [14,60]. A stylized simulation model serves as the ground truth, from which we extract data representing the observed data of the system. Fig. 3 visualizes the ground truth set-up using the following steps, where the steps are also shown in the Figure:

Step 1: Start with the input data for the ground truth, a supply chain graph with vertices and edges, $g^o = (V^o, E^o)$, and input parameters, X^o. The ground truth was presented in Section 3.1, and is detailed in Appendix A.1 and Appendix A.2.

Step 2: Use the input data to design and run a ground truth simulation model. From this model, we extract the ground truth data.

Step 3: Degrade the ground truth data on three dimensions of data sparseness (missing values, noise, and bias) following van Schilt et al. [14]. Missing values relate to the incompleteness of the available information, and these are created by removing records from the ground-truth data and setting values in the data to null or NaN (Not-a-Number). Noise relates to measurement issues for data and are created by replacing values v from the ground-truth data by $v+\epsilon$, where $\epsilon \sim N(0, \sigma^2)$. Bias indicates that certain values are more likely to show up in the measured data than others, e.g., because they are easier to measure. Bias is created by replacing ground truth data by repeated records with the same data that therefore show up more often and 'hide' some of the ground truth data. For the experiments, we design scenarios that suffer from all three dimensions of data sparseness, reflecting real-world data of supply chains. For each of the dimensions, we choose a low sparseness of 20%, a medium sparseness of 50%, or a high sparseness of 80%. Table 2 presents an overview of the ten scenarios used in this study.

Table 2
Scenarios for the applied percentages of data sparseness in step 3.

Scenarios	Bias	Noise	Missing values
No sparseness	0%	0%	0%
All Low	20%	20%	20%
All Medium	50%	50%	50%
All High	80%	80%	80%
Bias Low	20%	50%	50%
Bias High	80%	50%	50%
Noise Low	50%	20%	50%
Noise High	50%	80%	50%
Missing Low	50%	50%	20%
Missing High	50%	50%	80%

Step 4: From the input space, we select candidate input space parameters (a graph number for the structure, and parameter values for the other 32 input space parameters) as candidate solutions. Then, the simulation runs 10 replications with different seeds, and produces the simulated output, after which the objective value (average Manhattan distance to the sparse data over the 10 replications) is calculated. The average total transport cost and the density of the network are given as additional outputs, since this determines the location of the solution in the behavior space. The QD process continues until a stopping criterion is reached.

Step 5: Collect the set of optimal solutions resulting from the QD process. This set of optimal solutions, S, contains one optimal solution for each QD container in the behavior space, i.e., $s^*_{C,R} \in S$. Each solution is a combination of a graph and input parameters that assign the solution to a part of the behavior space grid, i.e., $s^*_{C,R} = (g^*_{C,R}, X^*_{C,R})$ where $g^*_{C,R} = (V^*_{C,R}, E^*_{C,R})$.

Step 6: Analyze the quality-of-fit by comparing the ground truth input and the set of solutions resulting from QD. While QD minimizes the gap between the simulated data and the sparse data, this does not necessarily mean that the set of solutions captures the ground truth. Hence, we determine the proximity of the set of solutions to the ground truth by assessing how often the ground truth is identified by QD across various unique seeds. In addition, we compare the solutions on various properties such as transport cost, density, objective value (Manhattan distance), the number of vertices, the number of edges, and the graph edit distance, i.e., the cheapest set of graph edit operations needed to transform the graph selected in the input space to the graph representing the ground truth [61]. We use an approximated greedy graph edit distance as described by Riesen et al. [62] for computational reasons.

4. Results

We present the results of calibrating the supply chain simulation model for each feature profile using the QD algorithm when varying the degree of data sparseness. First, we present the results of the convergence of the QD algorithm. Second, we analyze the extent to which QD can identify the ground truth across the seeds. Third, we combine the results of the runs with different seeds into a single QD front and examine the QD container where the ground truth is expected. Last, we evaluate the overall QD front.

In this section, we refer to the Quality Diversity results mapped into the behavior space as the *QD front*. Fig. 4 presents an example of the QD front of a scenario with 0% data sparseness. In this figure, we identify the ground truth values in the behavior space with a red dot. The behavior values of the ground truth model are $709.9 for transport cost and 0.0573 for density. In the discretized behavior space, this means the ground truth fits in the QD container with values between $650 to $750 for transport cost and a density of 0.055 to 0.060. We refer to this container as the *ground truth container*.

To properly compare the different data sparseness scenarios, the output space, in terms of its Manhattan (L1) distance, is normalized using the minimum and the maximum objective values of the QD front. This means the objective value closest to the ground truth is 0, and the objective value the farthest away from the ground truth is 1. The direction of desirability is towards 0. We refer to the normalized objective value as normalized L1 distance. More information on the normalization can be found in Appendix A.4.

4.1. Convergence of QD algorithm

We evaluate the convergence of the QD algorithm for calibrating the structure and the parameters of the counterfeit PPE simulation model. For this, we focus on the scenario with 0% data sparseness. A measure for convergence is the quality of the QD front per seed by the average L1 distance between the QD containers and the ground truth. This measure indicates the magnitude of improvements from each function evaluation. A constant value means no substantial improvements are made to the front. Another measure for convergence is the diversity of the QD front per seed by the coverage, i.e., the percent of QD containers in the behavior space that contain a solution. We divide the behavior space into 100 QD containers, so a coverage of 0.94 means that 94 of the 100 QD containers contain a solution.

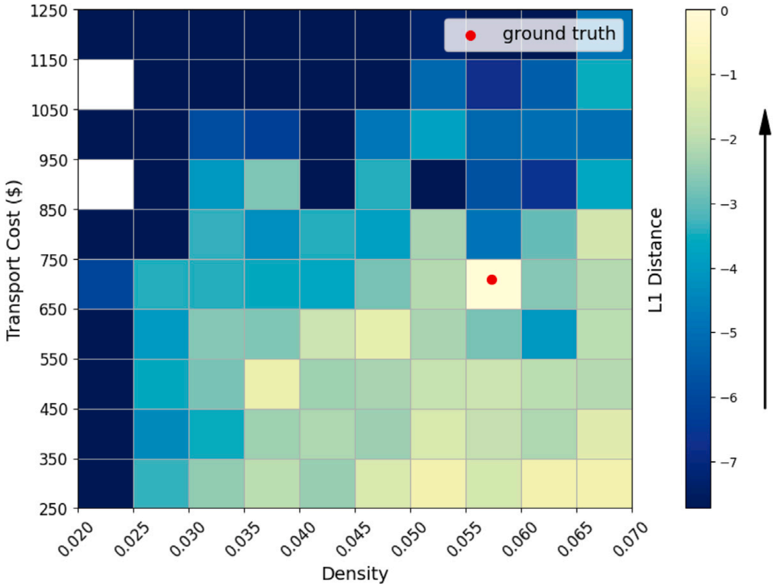
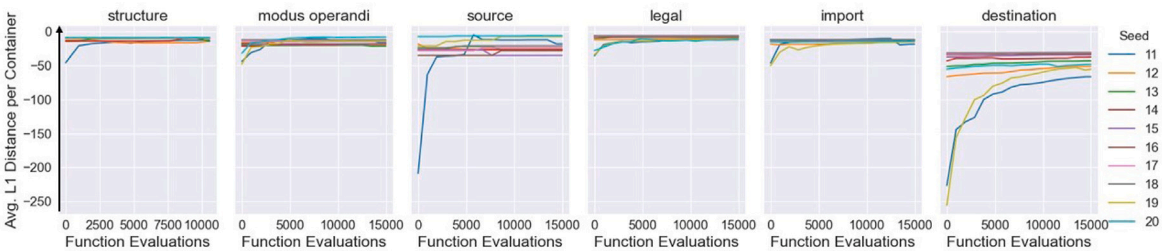
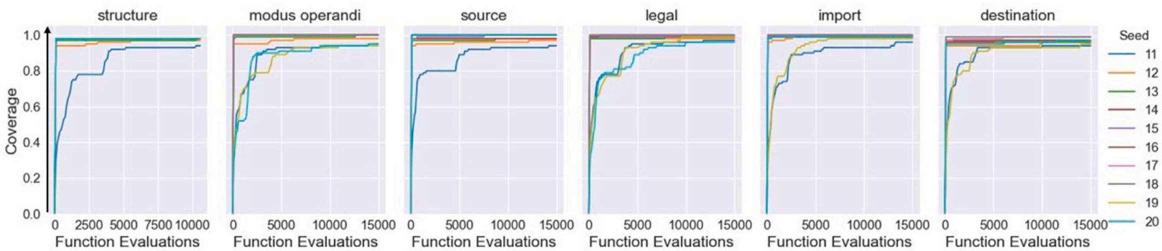


Fig. 4. Quality Diversity results of a scenario with 0% data sparseness.



(a) Quality Measured as Average L1 Distance Between the QD Containers and the Ground Truth per Feature Profile



(b) Density Measured as Coverage Percent

Fig. 5. Convergence on quality and diversity per seed at 0% of data sparseness.

Fig. 5 shows that the average L1 distance and the coverage are constant for most of the seeds for all the feature profiles. In general, the average L1 distance across the QD containers and the coverage stays relatively constant starting from the initial sample, except for seed 11 and seed 19. There is still much room for improvement for these seeds after initializing, but the average L1 distance and the coverage become closer in proximity to the other seeds after more function evaluations. Overall, Fig. 5(a) shows that calibrating for the profiles ‘source’ and ‘destination’ leads to the variation between the seeds on the average L1 distance across the QD containers. Fig. 5(b) shows that most seeds have a high coverage and, therefore, lead to a high degree of diversity.

In some experiments, the figures show that the average L1 distance over the QD containers decreases, whereas the aim is to increase towards 0. This can be explained by the diversity of the QD front in that experiment. For example, for the feature profile ‘source’ at seed 11, we see a peak of the average L1 distance at 5760 function evaluations of −4.9 with a coverage of 0.74. For the

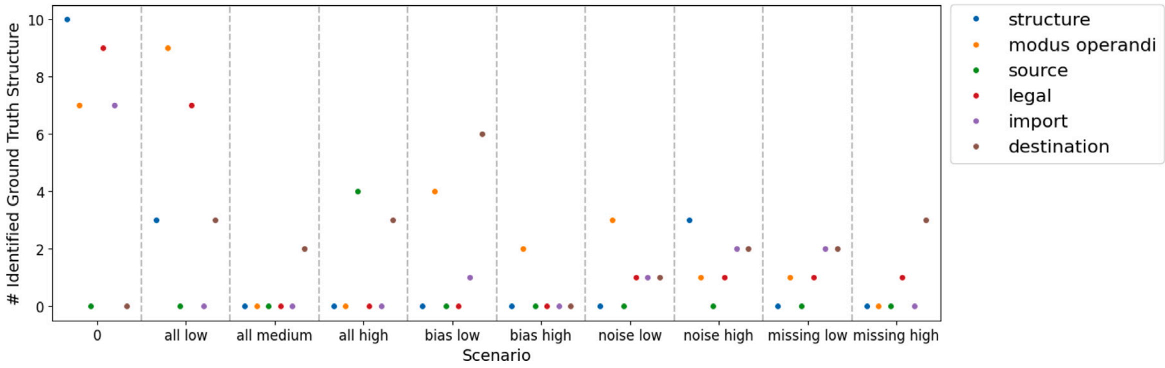


Fig. 6. Number of times the ground truth structure is identified per feature profile and per scenario across ten seeds.

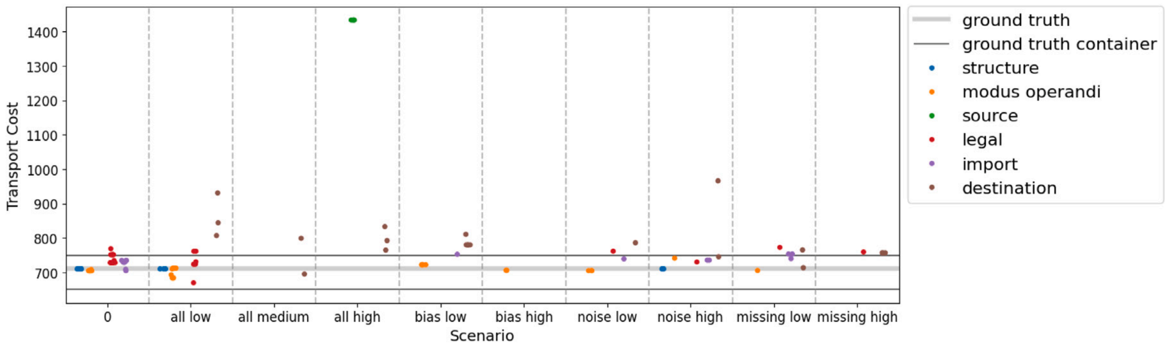


Fig. 7. Transport cost of the identified ground truth structures per feature profile and per scenario across ten seeds.

next iteration at 6720 function evaluations, the average L1 distance across the QD containers decreases to −11.8 with a coverage of 0.76. This means that although the average L1 distance decreases, the coverage increases.

4.2. Identifying the ground truth across seeds

Fig. 6 displays how many times the ground truth structure has been identified across the ten seeds for each feature profile and the various data sparseness scenarios. Overall, the ground truth structure is most often identified at 0% data sparseness, especially for the feature profile where only the structure is calibrated. Remarkably, for the profile 'source' and the profile 'destination', QD fails to identify the ground truth structure at 0% data sparseness. For the profile 'source', the ground truth structure is successfully identified only once for the scenario 'all high'. For the profile destination, the ground truth structure has been identified in scenarios with more data sparseness and, most frequently, in the case of the scenario with a low bias percentage. For the other feature profiles, the results generally show that more sparseness leads to a lower or similar identification of the ground truth structure. Apparently, there are other supply chain structures that are able to reproduce the data from the scenario better or equally well. Note that especially in case of high bias, the data presented to the QD algorithm contains 'false' information, not representative of the actual ground truth.

Examining the solutions that contain the ground truth structure, Fig. 7 shows that many of these solutions have a higher transport cost compared to the ground truth. This means that these solutions are placed in QD containers other than the ground truth container. Especially for the profiles legal and destination, the identified ground truth transport cost is often higher than the ground truth. This could mean that the QD calibrates the legal and destination parameters often too high compared to the ground truth such that the configuration results in a higher transport cost. For the source feature profile, we see that, in the only scenario where the ground truth structure has been identified, the configuration of the source parameters leads to extremely high transport costs. This is caused by relatively low interarrival time and relatively high warehouse consolidator times. Thus, the behavior of this specific solution does not align closely with the ground truth in terms of transport cost. See Appendix A.5 for the graphs on the parameter values.

Zooming in on the QD container in the behavior space where the ground truth is expected, Fig. 8 displays the optimal supply chain configurations for each seed over the various scenarios and feature profiles in this specific QD container. When combining the results of the seeds into a single QD front, the solution with the lowest L1 distance is chosen. In the figure, we see that the solution containing the ground truth structure is optimal for the majority of the profiles at 0% data sparseness, and the scenario 'all low'.

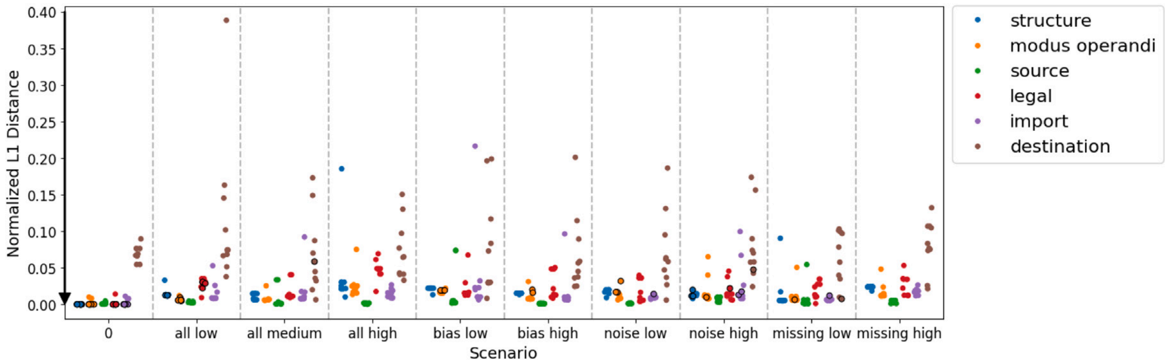


Fig. 8. Normalized L1 distance for the ground truth container per feature profile and per scenario across ten seeds. The solutions containing the ground truth structure have a black outer edge. The arrow represents the direction of desirability.

However, for the other scenarios for data sparseness, the solution containing the ground truth frequently fails to be optimal across the seeds and, consequently, does not appear in the single QD front.

4.3. Analyzing the ground truth container of QD front

The solutions of the seeds are combined to create a single QD front for each feature profile in each scenario. Similar to the QD algorithm, the best optimal solution for each QD container across the seeds is included in the QD front. We analyze the QD container where the ground truth model fits to assess the quality-of-fit of the QD front for each feature profile in each scenario.

Fig. 9 shows the *density*, i.e., the ratio of the number of edges to the possible number of edges in the graph, and the *graph edit distance*, i.e., the cheapest set of graph edit operations to transform the graph to the ground truth. The figure shows that the ground truth is most often identified for the scenario with 0% data sparseness and the scenario with all dimensions having a low data sparseness. For the other scenarios, the optimal solutions found in the ground truth container have a higher density than the ground truth graph (Fig. 9(a)). The graph edit distance of solutions that did not result in the ground truth structure is mostly between 276 and 621 edit operations (Fig. 9(b)). This result implies that when the data quality decreases, more complex supply chain graphs fit this data better. In other words, more complex models fit the data with noise, bias, and missing values better than the 'simpler' ground truth model.

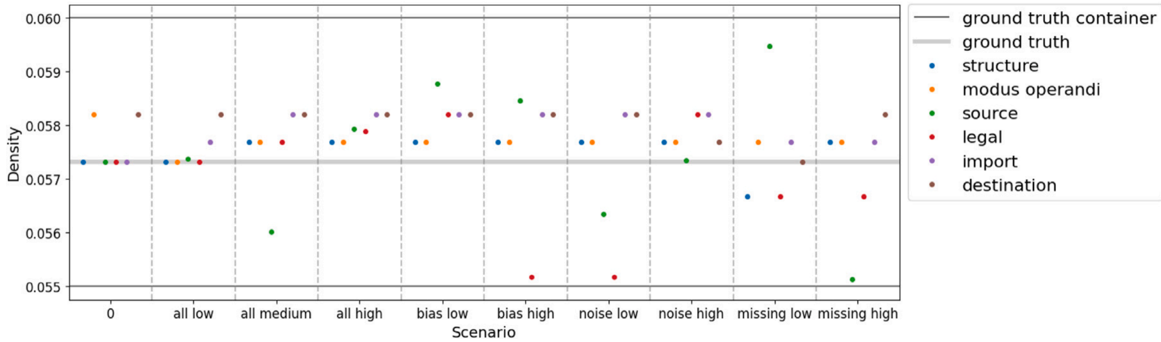
Fig. 9 shows the two graph structures that are most often identified as optimal across most scenarios. The graph structure with a density of 0.0577 and a graph edit distance of 621 is often identified as optimal for the feature profiles 'structure' and 'modus operandi'. The graph structure with a density of 0.0582 and a graph edit distance of 276 is often identified as optimal for the profiles 'import' and 'destination'. An exception is the profile 'source' that did not identify either of the two optimal graph structures for any scenario. This profile has identified solutions with a relatively high or relatively low graph edit distance compared to the ground truth. The solutions identified for this profile have the highest graph edit distance, with a range between 362 and 873.

Table 3 presents the ground truth structure and the graph structures of the two most often identified optimal solutions. The graph with a density closer to the ground truth (1) has a higher graph edit distance and a higher number of vertices and edges than the graph with a higher density (2). The figures of the graph structures also show that graph (1) has many more transit ports, import ports, distributors, and hospitals than graph (2). Interestingly, all three structures have one supplier, one or two manufacturers, and one warehouse consolidator. Additionally, all graphs have a combination of sea and air transport, with an overlap of ports in the graphs, such as Hong Kong, Amsterdam, Shanghai, and Singapore. Notably, the three graphs have the airport in Boston as an import port. Additional figures on the transport costs, the number of vertices, and the number of edges for each solution in the ground truth container in the QD front can be found in Appendix A.6.

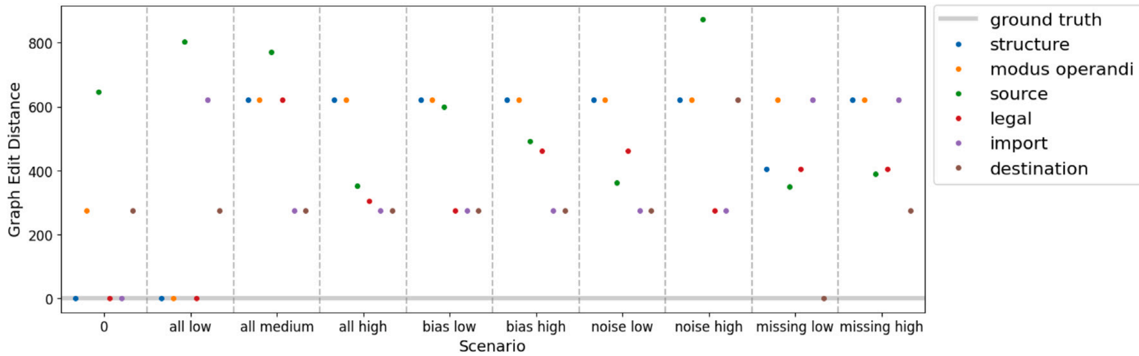
4.4. Analyzing the overall QD front

We analyze the overall QD front per feature profile and per scenario on diversity and quality. Regarding diversity, the QD front for each scenario and each profile has a relatively high coverage when merging the ten unique seeds (see Fig. 5). For the single QD front, the coverage is between 0.94 and 1.0.

Fig. 10 shows the distribution of quality of the solutions in the QD front using the number of occurrences (count) of the normalized L1 distance. A histogram with a binwidth of 0.1 is used to determine the counts, meaning there are 10 bins in total. We refer to each bin by using the minimum value of that particular bin, e.g., the bin of 0.1 to 0.2 is referred to as the value 0.1. A high count suggests that this value is more frequently observed in the QD front. In general, we see that each feature profile has a high count around a normalized L1 distance of 0.0, meaning that most solutions in the QD front are relatively close to the ground truth. The distribution is skewed towards the left. Most feature profiles have a normalized L1 distance of 0.4 to 0.8, with another small peak around the bin of 0.9.



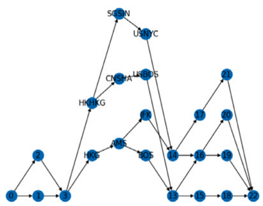
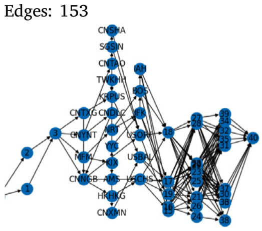
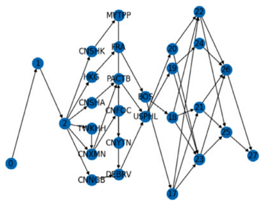
(a) Density.



(b) Graph Edit Distance.

Fig. 9. Characteristics of the solutions in the ground truth container in the Quality Diversity front per scenario and per feature profile.

Table 3
Characteristics of the graph structure of the ground truth and the two most often identified solutions (1)–(2) in the ground truth container of the Quality Diversity front. Each step in the horizontal line of the structure plot represents a set of actors, going from left to right: supplier, manufacturer, warehouse consolidator, export port, transit port, import port, warehouse distributor, distributor, and hospital.

Ground truth	(1)	(2)
Density: 0.0573 Graph Edit Distance: 0 Vertices: 23 Edges: 29 	Density: 0.0577 Graph Edit Distance: 621 Vertices: 52 Edges: 153 	Density: 0.0582 Graph Edit Distance: 276 Vertices: 28 Edges: 44 

When looking at the modus operandi and legal feature profiles, the figure shows two different directions of counts across the data sparseness scenarios. One has a higher number of occurrences around a normalized L1 distance of 0.2, whereas the other has no occurrences around 0.2. The results show that the scenarios that have a high number of occurrences in the modus operandi profile tend to have a low number of occurrences in the legal profile. For example, the scenarios with 0% data sparseness and ‘all high’ have no occurrences at 0.2 of normalized L1 distance for the modus operandi profile. In contrast, the scenario of 0% data sparseness and ‘all high’ have a count of 10 at a normalized L1 distance of 0.2.

For the other feature profiles (structure, source, import, and destination), a difference in the number of occurrences between the scenarios is shown around the value 0. For structure, source, and import, we see that all scenarios have the highest number of occurrences around the value 0.1 for the normalized L1 distance. Generally, the scenario ‘bias low’ generally has the lowest count.

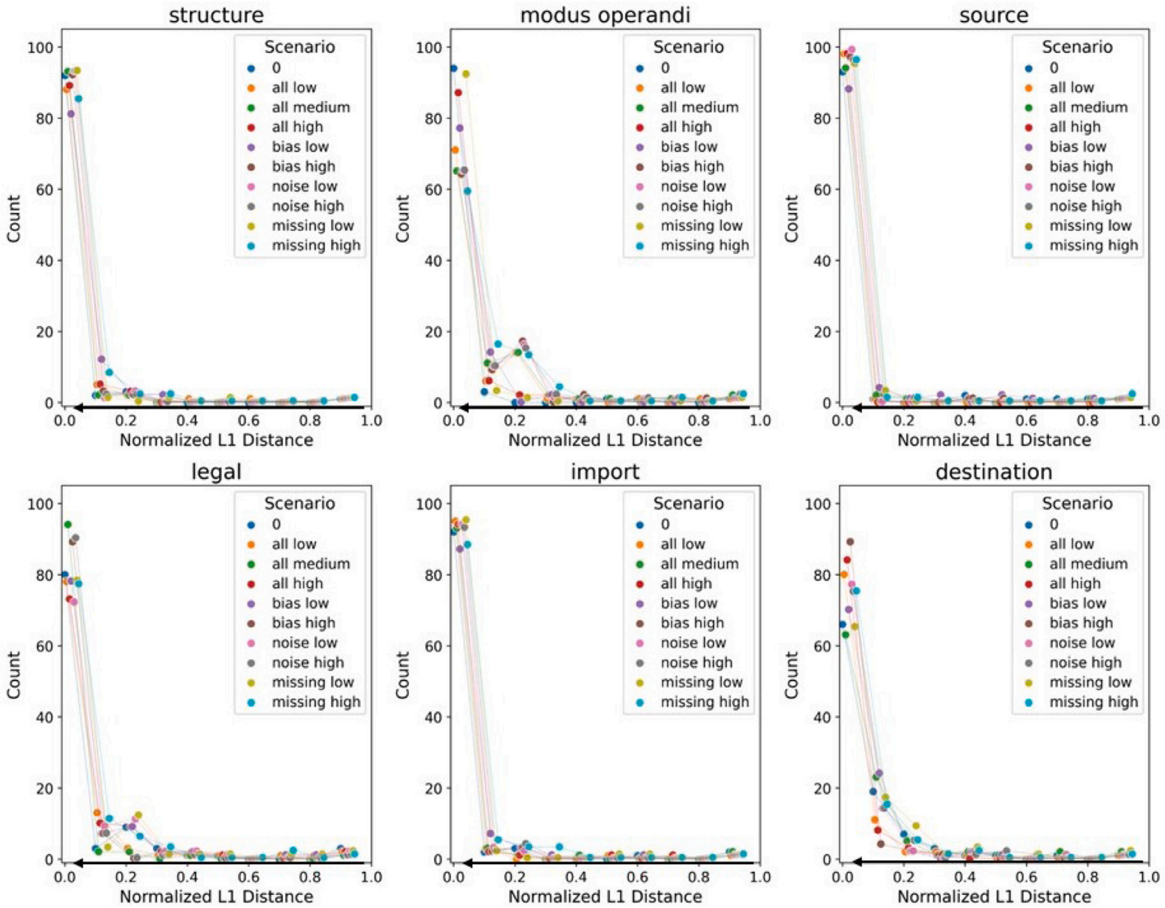


Fig. 10. Number of occurrences (Count) of the normalized L1 distance of the solutions in the Quality Diversity front. A histogram with a binwidth of 0.1 is used for the counts, meaning a total of 10 bins between 0 and 1. The points are the count of one particular bin, e.g., 80 occurrences between 0.0 and 0.1, and are plotted at the starting value of the bin. For visualization purposes, we added some jitter to limit the overlap in the data points.

For destination, we see a high dispersion in the number of occurrences between the scenarios for the normalized L1 distance of 0, where the scenario 'bias high' has the highest count and the scenario 'all medium' has the lowest.

For the source, legal, import, and destination feature profiles, the scenario with 0% data sparseness is on the relatively low segment for the number of occurrences at the value 0, with several occurrences for a higher normalized L1 distance. This could suggest that the scenarios with more sparseness lead to more solutions with a lower normalized L1 distance. Comparatively, the scenario with 0% data sparseness has a relatively high number of occurrences for the value 0, and smaller counts for a higher normalized distance for the profiles 'structure' and 'modus operandi'. This suggests that the scenarios with more data sparseness lead to more solutions with a higher normalized L1 distance.

5. Discussion

This section reflects on the results of the Quality Diversity (QD) algorithm and discusses the limitations of the research.

5.1. Reflection on using a quality diversity algorithm

The QD algorithm successfully shows its feasibility for calibrating a supply chain simulation model. A key notion for using QD in this field is its sensitivity to the initialization of the seeds of the algorithm, which determines the initial sample and the randomness for selecting candidate solutions. The convergence results show that for some seeds, there is still much room for improvement in terms of the average Manhattan distance of the solutions and the coverage, whereas other seeds instantly reach a satisfactory level of quality and diversity. An explanation for this is that, in a highly rugged fitness landscape typical for discrete event simulation [63], QD needs to perform additional iterations to reach convergence when the initial sample is chosen poorly. Thus, when using QD to calibrate a discrete event simulation model, it is crucial to use various seeds.

In terms of diversity, the results demonstrate that QD fills at least 96% of the QD containers in the behavior space for calibrating this counterfeit PPE supply chain model across all input space features and scenarios for data sparseness. Although the main limitation of QD is that it does not guarantee to fill every QD container in the discretized behavior space [20,21,50], QD successfully reaches a high coverage for calibrating this simulation model. Nevertheless, it is necessary to consider the trade-off between the *diversity* enforced by the algorithm, e.g., solutions are "all over the place", and the *quality* of the QD front, e.g., solutions that are highly optimal.

In terms of the quality-of-fit for calibrating the structure, the results show that QD is able to identify the ground truth for most feature profiles and various scenarios for data sparseness across the seeds. The ground truth structure is, as expected, most frequently identified at 0% data sparseness. In the case of more data sparseness, the solution containing the ground truth structure often has a higher normalized L1 distance and is not the most optimal solution across the seeds. Thus, solutions with another graph structure better fit the sparse data than the ground truth. These solutions have high graph edit distances, and more vertices and edges than the ground truth. Simulation models with more complex structures often reproduce the sparse data better than those that slightly differ from the ground truth, indicating a risk of overfitting. However, the simulation models with more complex structures do not necessarily lead to "wrong" results since they do generate similar results as provided by the sparse data.

Regarding the quality-of-fit for calibrating the actors' parameters, the results demonstrate that the parameters of the source and destination actors seem to have the most impact on the simulation outcomes and the resulting QD front. For the source-related features, QD has difficulty fitting the parameters and the structure, even with 0% data sparseness. For the destination-related features, the ground truth structure is more often identified when data is sparse, but the identified actors' parameters result in high transport costs, leading to a different container in the QD front. For the other input space features, the actors' parameters are not necessarily similar to the ground truth, but the QD front is not highly sensitive to them. Even though some input space features have a higher number of parameters to calibrate, the individual parameters of each feature have less impact on the simulation model outcomes.

However, the number of the parameters does appear to have an impact on the QD algorithm. Specifically, for the feature profiles with the fewest parameters to calibrate, 1 for structure and 4 for modus operandi, the scenario with 0% data sparseness shows a high number of occurrences of a normalized L1 distance equal to zero in the overall QD front. This can be explained by the nature of the QD algorithm CMA-ME, which uses a covariance matrix [51]. With fewer parameters, the covariance structure is simpler, enhancing the exploration of solutions with 0% data sparseness close to the ground truth. For calibrating a simulation model with more than 5 parameters, the results demonstrate that more data sparseness could lead to more solutions that are coherent with the sparse data. In a sense, data sparseness, such as missing values, creates more degrees of freedom, for which the QD algorithm can find fitting solutions.

5.2. Limitations

The QD method may become less effective when attempting to calibrate many input parameters. Even though we have chosen a subset of features and bounded the input space parameters to limit the input space for the QD algorithm, it is still quite challenging to calibrate the 33 parameters in the input space using a simulation model. Especially the structure parameter has many degrees of freedom; in our case, there was a choice between 40,000 different graphs. To combine each of the graph options with value combinations of the other 32 input space parameters, millions of evaluations might be needed. Each evaluation involves 10 simulation replications. This has impact on the computational complexity of calibrating a simulation model such as the number of function evaluations required. QD papers show that for input spaces with few parameters, the covariance matrix is simpler, enhancing the exploration of solutions closer to the ground truth. Further research needs to explore the limits of the usability of the QD method in relation to the number of parameters.

Using a System Entity Structure, we have created a large set of different possible configurations of the supply chain. However, given the many degrees of freedom within a supply chain structure (number of actors, possible links, different travel distances etc.), the set might not capture all combinations needed for the fit.

The choice of the behavior space is crucial for applying the QD results in real life. In this study, we only focused on the transport cost and on network vulnerability, whereas other factors, such as the cost of production, the market value of the goods, and the trust between actors, also play a role in supply chains. For further research, it would be interesting to expand the behavior space to incorporate other defining properties of a supply chain.

Finally, the fact that only one case study was used is a clear limitation to the generalizability of the research. Since the simulation calibration approach was independent of the specific supply chain in the case study, the method would stay the same when widening the application area. Of course, the underlying simulation model would change. It would be interesting to evaluate whether the results still hold for other types of supply chains or systems.

6. Conclusion

This research examines the feasibility of using the Quality Diversity (QD) algorithm for generating a diverse ensemble of supply chain simulation models when the available data is sparse. For this, we use a case study of a counterfeit PPE supply chain as the ground truth, extract data from the ground truth, and vary the degree of data sparseness. We assess whether QD can identify the ground truth among the diverse set of solutions, in the case of structural and parametric uncertainty.

Our analysis demonstrates that QD is able to generate a diverse ensemble of supply chain simulation models. Due to the algorithms' sensitivity to seed initialization, it is crucial to use a number of seeds. When the data sparseness increases, simulation models with more complex structures, i.e., more vertices and edges than the ground truth, tend to describe and reproduce the sparse data better. These complex structures are not necessarily "wrong", as they show overlap with the ground truth in parts of the supply chain, but they still overfit the sparse data. For parametric uncertainty, the impact of the parameters on the simulation model outcomes, rather than the number of parameters, influences the quality-of-fit of the solutions of the QD algorithm for identifying the ground truth.

In the case of structural and parametric uncertainty, QD clearly has two major advantages over classical calibration methods: it can fit model structure as well as the parameter values, and it provides a broad set of diverse solutions rather than a single optimal fit.

Further research should focus on applying the method to other systems than supply chains. Furthermore, it would be important to investigate the effects of increasing the number of dimensions of the behavior space, both on the diversity of the solutions generated and on the understandability of the results.

CRediT authorship contribution statement

Isabelle M. van Schilt: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Project administration. **Jan H. Kwakkel:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Jelte P. Mense:** Conceptualization, Writing – original draft, Resources, Visualization, Validation, Supervision. **Alexander Verbraeck:** Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing, Validation, Supervision.

Funding

This research project was fully funded by the National Police Artificial Intelligence Lab (the ICAI NPAI Lab) of the Netherlands.

Research data

The code and data used for this project can be found at

- https://doi.org/10.4121/48bdc725-a1db-4b87-8508-aee808713957.v1
  Python generator of the structures for the simulation model of the stylized PPE supply chain.
- https://doi.org/10.4121/766f4e89-fa03-47c6-a9f2-fa41f241984b.v1
  Python code and data for the application of the QD algorithm.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.simpat.2025.103216.

Data availability

The paper lists two doi's for the data.

References

[1] I.A. Omar, M. Debe, R. Jayaraman, K. Salah, M. Omar, J. Arshad, Blockchain-based supply chain traceability for COVID-19 personal protective equipment, Comput. Ind. Eng. (2022) 107995, http://dx.doi.org/10.1016/j.cie.2022.107995.

[2] L. Hashemi, E. Huang, L. Shelley, Counterfeit PPE: Substandard respirators and their entry into supply chains in major cities, Urban Crime. Int. J. 3 (2) (2022) 74–109, http://dx.doi.org/10.26250/heal.panteion.uc.v3i2.290.

[3] M. Ippolito, C. Gregoretti, A. Cortegiani, P. Iozzo, Counterfeit filtering facepiece respirators are posing an additional risk to health care workers during COVID-19 pandemic, Am. J. Infect. Control 48 (7) (2020) 853, http://dx.doi.org/10.1016/j.ajic.2020.04.020.

[4] L. Hashemi, C.C. Jeng, A. Mohiuddin, E. Huang, L. Shelley, Simulating Counterfeit Personal Protective Equipment (PPE) Supply Chains During COVID-19, in: B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C. Corlu, L. Lee, E. Chew, T. Roeder, P. Lendermann (Eds.), Proceedings of the 2022 Winter Simulation Conference, Institute of Electrical and Electronics Engineers, Inc, Singapore, 2023, pp. 522–532, http://dx.doi.org/10.1109/WSC57314.2022.10015398.

[5] I.M. van Schilt, J. Kwakkel, J.P. Mense, A. Verbraeck, Calibrating simulation models with sparse data: Counterfeit supply chains during Covid-19, in: B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C. Corlu, L. Lee, E. Chew, T. Roeder, P. Lendermann (Eds.), Proceedings of the 2022 Winter Simulation Conference, Institute of Electrical and Electronics Engineers, Inc, Singapore, 2023, pp. 496–507, http://dx.doi.org/10.1109/WSC57314.2022.10015241.

[6] J. Banks, Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice, John Wiley & Sons, 1998.

[7] B.P. Zeigler, A. Muzy, E. Kofman, Theory of Modeling and Simulation: Discrete Event & Iterative System Computational Foundations, third ed., Academic Press, FL, USA, 2018, http://dx.doi.org/10.1016/C2016-0-03987-6.

[8] A. Schmitt, M. Singh, Quantifying supply chain disruption risk using Monte Carlo and discrete-event simulation, in: M. Rossetti, R.R. Hill, B. Johansson (Eds.), Proceedings of the 2009 Winter Simulation Conference, IEEE, Austin, TX, USA, 2010, pp. 1237–1248, http://dx.doi.org/10.1109/WSC.2009.5429561.

[9] N.R. Magliocca, A.N. Price, P.C. Mitchell, K.M. Curtin, M. Hudnall, K. McSweeney, Coupling Agent-Based Simulation and Spatial Optimization Models to Understand Spatially Complex and Co-Evolutionary Behavior of Cocaine Trafficking Networks and Counterdrug Interdiction, Inst. Ind. Syst. Eng. Trans. (2022) 1–14, http://dx.doi.org/10.1080/24725854.2022.212399.

[10] M.R. Wigan, The fitting, calibration, and validation of simulation models, Simulation 18 (5) (1972) 188–192, http://dx.doi.org/10.1177/003754977201800506.

[11] T.I. Ören, Concepts and criteria to assess acceptability of simulation studies: A frame of reference, Commun. Assoc. Comput. Mach. 24 (4) (1981) 180–189, http://dx.doi.org/10.1145/358598.358605.

[12] M. Hofmann, On the complexity of parameter calibration in simulation models, J. Déf. Model. Simul. 2 (4) (2005) 217–226, http://dx.doi.org/10.1177/154851290500200405.

[13] S. Somapa, M. Cools, W. Dullaert, Characterizing supply chain visibility - a literature review, Int. J. Logist. Manag. 29 (1) (2018) 308–339, http://dx.doi.org/10.1108/IJLM-06-2016-0150.

[14] I.M. van Schilt, J.H. Kwakkel, J.P. Mense, A. Verbraeck, Dimensions of data sparseness and their effect on supply chain visibility, Comput. Ind. Eng. 191 (2024) 110108, http://dx.doi.org/10.1016/j.cie.2024.110108.

[15] A. Ficara, L. Cavallaro, F. Curreri, G. Fiumara, P. De Meo, O. Bagdasar, W. Song, A. Liotta, Criminal Networks Analysis in Missing Data Scenarios Through Graph Distances, PLOS ONE 16 (8) (2021) e0255067, http://dx.doi.org/10.1371/journal.pone.0255067.

[16] S.D. Mitchell, Integrative pluralism, Biology Philos. 17 (2002) 55–70, http://dx.doi.org/10.1023/A:1012990030867.

[17] W. Veit, Model pluralism, Philos. Soc. Sci. 50 (2) (2020) 91–114, http://dx.doi.org/10.1177/00483931198948.

[18] S.E. Page, The Model Thinker:what You Need to Know to Make Data Work for You, second ed., Hachette Book Group, New York, NY, USA, 2021.

[19] J.M. Durán, N. Formanek, Grounds for trust: Essential epistemic opacity and computational reliabilism, Minds Mach. 28 (2018) 645–666, http://dx.doi.org/10.1007/s11023-018-9481-6.

[20] J.-B. Mouret, J. Clune, Illuminating search spaces by mapping elites, 2015, http://dx.doi.org/10.48550/arXiv.1504.04909, arXiv preprint arXiv:1504.04909.

[21] K. Chatzilygeroudis, A. Cully, V. Vassiliades, J.-B. Mouret, Quality-diversity optimization: a novel branch of stochastic optimization, in: P. Pardalos, V. Rasskazova, M. Vrahatis (Eds.), Black Box Optimization, Machine Learning, and No-Free Lunch Theorems, Springer Optimization and Its Applications, Springer, Cham, 2021, pp. 109–135, http://dx.doi.org/10.1007/978-3-030-66515-9_4.

[22] J.K. Pugh, L.B. Soros, P.A. Szerlip, K.O. Stanley, Confronting the challenge of quality diversity, in: S. Silva (Ed.), Proceedings of the Genetic and Evolutionary Computation Conference, Association for Computing Machinery, New York, NY, USA, 2015, pp. 967–974, http://dx.doi.org/10.1145/2739480.2754664.

[23] B. Lim, M. Allard, L. Grillotti, A. Cully, Accelerated quality-diversity for robotics through massive parallelism, in: Workshop on Agent Learning in Open-Endedness, International Conference of Learning Representations, 2022, Online. URL https://openreview.net/forum?id=SUqU4mPbUZ9.

[24] B. Tjanaka, M.C. Fontaine, D.H. Lee, Y. Zhang, N.R. Balam, N. Dennler, S.S. Garlanka, N.D. Klapsis, S. Nikolaidis, Pyribs: A bare-bones python library for quality diversity optimization, 2023, http://dx.doi.org/10.48550/arXiv.2303.00191, arXiv preprint arXiv:2303.00191.

[25] T. Pierrot, G. Richard, K. Beguir, A. Cully, Multi-objective quality diversity optimization, in: Proceedings of the Genetic and Evolutionary Computation Conference, Association for Computing Machinery, New York, NY, USA, 2022, pp. 139–147, http://dx.doi.org/10.1145/3512290.3528823.

[26] L. Schneider, F. Pfisterer, J. Thomas, B. Bischl, A collection of quality diversity optimization problems derived from hyperparameter optimization of machine learning models, in: J.E. Fieldsend (Ed.), Proceedings of the Genetic and Evolutionary Computation Conference Companion, Boston, MA, USA, 2022, pp. 2136–2142, http://dx.doi.org/10.1145/3520304.3534003.

[27] P. Kent, J. Branke, Bayesian quality diversity search with interactive illumination, in: L. Paquete (Ed.), Proceedings of the Genetic and Evolutionary Computation Conference, Lisbon, Portugal, 2023, pp. 1019–1026, http://dx.doi.org/10.1145/3583131.3590486.

[28] V. Srikrishnan, K. Keller, Small increases in agent-based model complexity can result in large increases in required calibration data, Environ. Model. Softw. 138 (2021) 104978, http://dx.doi.org/10.1016/j.envsoft.2021.104978.

[29] L. Vanbrabant, N. Martin, K. Ramaekers, K. Braekers, Quality of input data in emergency department simulations: Framework and assessment techniques, Simul. Model. Pr. Theory 91 (2019) 83–101, http://dx.doi.org/10.1016/j.simpat.2018.12.002.

[30] R. Anzoom, R. Nagi, C. Vogiatzis, A Review of Research in Illicit Supply-Chain Networks and New Directions to Thwart Them, Inst. Ind. Syst. Eng. Trans. 54 (2) (2021) 134–158, http://dx.doi.org/10.1080/24725854.2021.1939466.

[31] Z. Liu, D. Rexachs, F. Epelde, E. Luque, A simulation and optimization based method for calibrating agent-based emergency department models under data scarcity, Comput. Ind. Eng. 103 (2017) 300–309, http://dx.doi.org/10.1016/j.cie.2016.11.036.

[32] A. De Santis, T. Giovannelli, S. Lucidi, M. Messedaglia, M. Roma, A simulation-based optimization approach for the calibration of a discrete event simulation model of an emergency department, Ann. Oper. Res. (2022) 1–30, http://dx.doi.org/10.1007/s10479-021-04382-9.

[33] L. de Groot, A. Hübl, Developing a Calibrated Discrete Event Simulation Model of Shops of a Dutch Phone and Subscription Retailer During COVID-19 to Evaluate Shift Plans to Reduce Waiting Times, in: S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, M. Loper (Eds.), Proceedings of the 2021 Winter Simulation Conference, Institute of Electrical and Electronics Engineers, Inc., Phoenix, Arizona, 2021, pp. 1–12, http://dx.doi.org/10.1109/WSC52266.2021.9715306.

[34] I.M. van Schilt, J.H. Kwakkel, J.P. Mense, A. Verbraeck, Identifying the structure of illicit supply chains with sparse data: A simulation model calibration approach, Adv. Eng. Inform. 62 (2024) 102926, http://dx.doi.org/10.1016/j.aei.2024.102926, URL https://www.sciencedirect.com/science/article/pii/S1474034624005779.

[35] M. Weisberg, Simulation and Similarity: Using Models to Understand the World, Oxford University Press, NY, USA, 2012.

[36] A. Tolk, E.H. Page, V.V. Graciano Neto, P. Weirich, N. Formanek, J.M. Durán, J.F. Santucci, S. Mittal, Philosophy and modeling and simulation, in: T. Ören, B.P. Zeigler, A. Tolk (Eds.), Body of Knowledge for Modeling and Simulation: A Handbook By the Society for Modeling and Simulation International, Springer, 2023, pp. 383–412, http://dx.doi.org/10.1007/978-3-031-11085-6_16.

[37] V.A.W.J. Marchau, W.E. Walker, P.J.T.M. Bloemen, S.W. Popper, Introduction, in: V.A.W.J. Marchau, W.E. Walker, P.J.T.M. Bloemen, S.W. Popper (Eds.), Decision Making under Deep Uncertainty: From Theory To Practice, Springer, Cham, Germany, 2019, pp. 1–20, http://dx.doi.org/10.1007/978-3-030-05252-2_1.

[38] E.L. Thompson, L.A. Smith, Escape from model-land, Econ.: Open-Access, Open-Assessment E-J. 13 (40) (2019) 1–15, http://dx.doi.org/10.5018/economics-ejournal.ja.2019-40.

[39] M.C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, J. R. Stat. Soc. Ser. B Stat. Methodol. 63 (3) (2001) 425–464, http://dx.doi.org/10.1111/1467-9868.00294.

[40] G. Evensen, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res.: Ocean. 99 (C5) (1994) 10143–10162, http://dx.doi.org/10.1029/94JC00572.

[41] G. Evensen, Data Assimilation: The Ensemble Kalman Filter, second ed., Springer, 2009, http://dx.doi.org/10.1007/978-3-642-03711-5.

[42] A. Tarantola, Inverse Problem Theory and Methods for Model Parameter Estimation, SIAM, 2005, http://dx.doi.org/10.1137/1.9780898717921.

[43] M.A. Beaumont, W. Zhang, D.J. Balding, Approximate Bayesian computation in population genetics, Genetics 162 (4) (2002) 2025–2035, http://dx.doi.org/10.1093/genetics/162.4.2025.

[44] M.A. Beaumont, Approximate Bayesian computation in evolution and ecology, Annu. Rev. Ecol. Evol. Syst. 41 (2010) 379–406, http://dx.doi.org/10.1146/annurev-ecolsys-102209-144621.

[45] A. Cully, Y. Demiris, Quality and diversity optimization: A unifying modular framework, Trans. Evol. Comput. 22 (2) (2017) 245–259, http://dx.doi.org/10.1109/TEVC.2017.2704781.

[46] D. Gravina, A. Liapis, G.N. Yannakakis, Quality diversity through surprise, Trans. Evol. Comput. 23 (4) (2018) 603–616, http://dx.doi.org/10.1109/TEVC.2018.2877215.

[47] L. Grillotti, A. Cully, Unsupervised behavior discovery with quality-diversity optimization, Trans. Evol. Comput. 26 (6) (2022) 1539–1552, http://dx.doi.org/10.1109/TEVC.2022.3159855.

[48] M. Fontaine, S. Nikolaidis, Differentiable quality diversity, Adv. Neural Inf. Process. Syst. 34 (2021) 10040–10052.

[49] V. Vassiliades, K. Chatzilygeroudis, J.-B. Mouret, Using Centroidal Voronoi Tessellations to scale up the multidimensional archive of phenotypic elites algorithm, Trans. Evol. Comput. 22 (4) (2017) 623–630, http://dx.doi.org/10.1109/TEVC.2017.2735550.

[50] J. Lehman, K.O. Stanley, Evolving a diversity of virtual creatures through novelty search and local competition, in: N. Krasnogor (Ed.), Proceedings of the Genetic and Evolutionary Computation Conference, Association for Computing Machinery, Dublin, Ireland, 2011, pp. 211–218, http://dx.doi.org/10.1145/2001576.2001606.

[51] M.C. Fontaine, J. Togelius, S. Nikolaidis, A.K. Hoover, Covariance matrix adaptation for the rapid illumination of behavior space, in: Proceedings of the Genetic and Evolutionary Computation Conference, Association for Computing Machinery, Cancún, Mexico, 2020, pp. 94–102, http://dx.doi.org/10.1145/3377930.3390232.

[52] L. Kuipers, Increasing Supply Chain Visibility with Limited Data Availability: Data Assimilation in Discrete Event Simulation, M.Sc. thesis, *Delft University of Technology*, 2021, URL https://resolver.tudelft.nl/uuid:5f68b82f-205e-4509-9a64-22082c46065f.

[53] P.H.M. Jacobs, The DSOL Simulation Suite (Ph.D. thesis), *Delft University of Technology*, 2005, http://dx.doi.org/10.4233/uuid:4c5586e2-85a8-4e02-9b50-7c6311ed1278.

[54] B.P. Zeigler, Multifacetted Modelling and Discrete Event Simulation, Academic Press Professional, Inc., San Diego, CA, USA, 1984.

[55] B.P. Zeigler, P.E. Hammonds, Modeling and Simulation-Based Data Engineering: Introducing Pragmatics into Ontologies for Net-Centric Information Exchange, Elsevier, USA, 2007.

[56] T. Snaphaan, T. van Ruitenburg, Financial crime scripting: an analytical method to generate, organise and systematise knowledge on the financial aspects of profit-driven crime, Eur. J. Crim. Policy Res. (2024) 1–21, http://dx.doi.org/10.1007/s10610-023-09571-9.

[57] J. Gao, C. Xie, C. Tao, Big data validation and quality assurance–Issuses, challenges, and needs, in: Symposium on Service-Oriented System Engineering, IEEE, Oxford, UK, 2016, pp. 433–441, http://dx.doi.org/10.1109/SOSE.2016.63.

[58] C.C. Aggarwal, A. Hinneburg, D.A. Keim, On the surprising behavior of distance metrics in high dimensional space, in: J. Van den Bussche, V. Vianu (Eds.), International Conference on Database Theory, Springer, London, UK, 2001, pp. 420–434, http://dx.doi.org/10.1007/3-540-44503-X_27.

[59] E.M. Mirkes, J. Allohibi, A. Gorban, Fractional norms and quasinorms do not help to overcome the curse of dimensionality, Entropy 22 (10) (2020) 1–31, http://dx.doi.org/10.1145/358598.358601.

[60] M. Khondoker, R. Dobson, C. Skirrow, A. Simmons, D. Stahl, A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies, Stat. Methods Med. Res. 25 (5) (2016) 1804–1823, http://dx.doi.org/10.1177/0962280213502437.

[61] Z. Abu-Aisheh, R. Raveaux, J.-Y. Ramel, P. Martineau, An exact graph edit distance algorithm for solving pattern recognition problems, in: A. Fred, M. De Marsico, M. Figueiredo (Eds.), 4th International Conference on Pattern Recognition Applications and Methods, Lisbon, Portugal, 2015, pp. 271–278, http://dx.doi.org/10.5220/0005209202710278.

[62] K. Riesen, M. Ferrer, R. Dornberger, H. Bunke, Greedy Graph Edit Distance, in: P. Perner (Ed.), International Conference of Machine Learning and Data Mining in Pattern Recognition, Springer, Hamburg, Germany, 2015, pp. 3–16, http://dx.doi.org/10.1007/978-3-319-21024-7_1.

[63] F. Azadivar, Simulation Optimization Methodologies, in: P. Farrington, H.B. Nembhard, D.T. Sturrock, G.W. Evans (Eds.), Proceedings of 1999 Winter Simulation Conference, Association for Computing Machinery, New York, NY, USA, 1999, pp. 93–100, http://dx.doi.org/10.1145/324138.324168.