Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C. G. Corlu, L. H. Lee, E. P. Chew, T. Roeder, and P. Lendermann, eds.

CALIBRATING SIMULATION MODELS WITH SPARSE DATA: COUNTERFEIT SUPPLY CHAINS DURING COVID-19

Isabelle M. van Schilt
Jan H. Kwakkel
Alexander Verbraeck

Faculty of Technology, Policy and Management
Delft University of Technology
Jaffalaan 5
Delft, 2628BX, NETHERLANDS

Jelte P. Mense

National Policelab AI
Utrecht University
Princetonplein 5
Utrecht, 3584CC, NETHERLANDS

ABSTRACT

COVID-19 related crimes like counterfeit Personal Protective Equipment (PPE) involve complex supply chains with partly unobservable behavior and sparse data, making it challenging to construct a reliable simulation model. Model calibration can help with this, as it is the process of tuning and estimating the model parameters with observed data of the system. A subset of model calibration techniques seems to be able to deal with sparse data in other fields: Genetic Algorithms and Bayesian Inference. However, it is unknown how these techniques perform when accurately calibrating simulation models with sparse data. This research analyzes the quality-of-fit of these two model calibration techniques for a counterfeit PPE simulation model given an increasing degree of data sparseness. The results demonstrate that these techniques are suitable for calibrating a linear supply chain model with randomly missing values. Further research should focus on other techniques, larger set of models, and structural uncertainty.

1 INTRODUCTION

During COVID-19, a rise in counterfeit Personal Protective Equipment (PPE) and related criminal activities was detected. Suddenly, there was a high worldwide demand for PPE such as face masks, particulate filter respirators, gloves, goggles, and glasses (Omar et al. 2022). Medical PPE for hospitals have stricter requirements, such as certification, than non-medical PPE. Certified PPE are more valuable than non-certified PPE, making it attractive for criminals to try and sell non-certified PPE as certified PPE. Detecting counterfeit PPE has been challenging since (1) COVID-19 is a new and unexpected phenomenon so there is little historical data, and (2) criminals generally try to share as little data as possible. Together, this makes it hard to get insight into criminal activities pertaining counterfeit PPE, making it a complex system.

Simulation is a way to get insight into complex systems, recognizing relations, and exploring future scenarios (Shannon 1998). In particular, the focus of this paper is on discrete event simulation for representing complex socio-technical systems (Schmitt and Singh 2009). A model can be conceptualized as consisting of variables and relations, and many variables need an initial value in the model to capture an initial state and behavior that is consistent with the state and behavior of the system. These initial values are called parameters; more specifically parameters of components of the model. Some of the parameter values might be observed directly, while others are unobservable and thus have to be tuned to match the behavior of the simulation model with its real world counterpart.

Model calibration can help with constructing a model close to the real world. It is the process of tuning and estimating the model parameters with observed data of the system to improve the similarity between the model and the system. The goal of model calibration is to find those parameter values for which the

behavior of the simulation model is as close as possible to the observed behavior of the real system by using real data.

In case of criminal activities in general, and in particular for counterfeit PPE, data is sparse. Criminals want to stay off the grid and generally do not voluntarily share information about their criminal activities. In case of COVID-19 related crimes, data sparseness is even more pronounced due to its novelty. This makes it even more challenging to calibrate models. In cases like this, model calibration should be able to handle sparse observed data. Data sparseness can be classified in three dimensions: (1) noise, (2) bias, and (3) missing values (Huang 2013; Hazen et al. 2014). This research focuses on one of the three dimensions of data sparseness, missing values. The goal of model calibration with sparse data is to find the most likely model configuration that matches the underlying system.

A subset of model calibration techniques seems to be able to handle sparse data in other fields. For example, Evolutionary Algorithms are widely applied for high-dimensional optimization problems where data often becomes sparse (Ren and Wu 2013). Bayesian Inference is often used for uncertainty analysis, and is one of the few techniques in machine learning that is able to handle sparse data sets (Vrugt and Beven 2018; Jalali et al. 2017). Data Assimilation is a promising technique for predicting simulation models in real-time with sparse data (Xie 2018; Kuipers 2021). However, it it yet unknown how these techniques perform for the calibration of simulation models given sparse data.

Therefore, this paper analyzes two model calibration techniques that are likely suitable for calibration in the case of sparse data. To test these techniques, a case study of a counterfeit PPE supply chain is used. We use a stylized discrete event simulation model of a counterfeit PPE supply chain as ground truth. We extract data from this model, systematically increase the degree of sparseness of the data, and assess the extent to which the selected model calibration techniques can still identify the underlying supply chain. We also test a commonly used model calibration technique as reference. This paper is the first step towards analyzing and comparing various model calibration techniques on simulation models of complex systems in the case of sparse data.

The paper is structured as follows. In Section 2, we discuss the current state-of-the-art literature on model calibration with sparse data, and select the model calibration techniques for this study. In Section 3, we explain the design of experiments used to test the selected model calibration techniques. In Section 4, we outline the simulation model of the case study, and present the results of the quality-of-fit of the selected model calibration techniques on the case study given an increasing degree of data sparseness. In Section 5, we discuss our results. In Section 6, we conclude our study, and provide some directions of further research.

2 MODEL CALIBRATION TECHNIQUES

Calibration of simulation models is defined as finding values for parameters of the model by using real data until there is a "good" agreement, i.e., as close as possible, between the model data and the observed data over a given time interval (Wigan 1972; Ören 1981; Hofmann 2005). Optimization techniques are commonly used for model calibration as the objective is to minimize the difference between the model data and the observed data (Liu et al. 2017).

2.1 Related Work

Malleson (2014) discusses the calibration of simulation models in the field of criminology. The author focuses on the goodness-of-fit in spatial structures. He presents three computer algorithms that help with exploring the parameter space: (1) Hill Climbing, (2) Simulated Annealing, and (3) Genetic Algorithms. Malleson (2014) emphasizes the need for gathering reliable observed data from the criminal system as this is not present yet. He notes that the calibrated model would not represent the real system when data is sparse. In our study, we do not focus on gathering this data but we focus on how to present the real system using model calibration given sparse observed data.

Liu et al. (2017) are one of the first to explicitly addresses calibration of a simulation model under data sparseness. They propose a simulation-optimization approach to automatically calibrate a simulation model with sparse data. They formulate the problem as a series of local minimum search problems. An agent-based model of an emergency department is used as case study. Following from this, De Santis et al. (2022) focus on calibration of a discrete event simulation model under data sparseness. They use the observable values from the target system for finding values of the simulation model on the level of model parameters, e.g., the time difference between known time stamps. de Groot and Hübl (2021) use calibration as a form of validation. In their case, validation of the simulation model is difficult due to the sparseness of data. They manually adjust parameters and behavior of the model to increase validity.

The main differences between the related work and our research are that (a) we compare various optimization techniques in the case of data sparseness instead of selecting one, and (b) we do not assume that one calibration technique works best for all types of sparse data.

2.2 Selected Techniques for Model Calibration with Sparse Data

We select a commonly used model calibration technique as reference technique: an exact solver using Powell's Method. As a first attempt to analyze the performance of techniques that seem to be able to deal with sparse data for calibrating simulation models, we select two model calibration techniques: Genetic Algorithms and a Markov Chain Monte Carlo sampling approximate Bayesian computation. The following sections describe these model calibration techniques in more detail.

2.2.1 Powell's Method

Exact solvers calibrate a model through exact mathematical optimization that guarantees to find (local or global) optimal solutions during model calibration (Puchinger and Raidl 2005). A commonly used exact algorithm for calibrating simulation models is Powell's Method (Liu et al. 2017). In a rugged high-dimensional fitness landscape typical for discrete event simulations, Powell's Method might be one of the best techniques for calibrating due to its fast search speed (Zhong and Cai 2015). Powell's Method is a gradient-free minimization algorithm using a repeated line search introduced by Powell (1964). In more detail, the algorithm selects a starting point and draws two different lines as search directions. On one of these lines, the algorithm performs an one-dimensional optimization to find a new optimal point. From this point on, an one-dimensional optimization is performed on the other line representing the different search direction. With these optimal points, a conjugate search direction is drawn where also an one-dimensional optimization is performed. These steps are repeated until the algorithm finds the optimal solution or when stopping criteria are reached (Vassiliadis and Conejeros 2009). In this research, the number of iterations and functions evaluations are used as stopping criteria.

2.2.2 Genetic Algorithm

Evolutionary algorithms calibrate a model through population-based, i.e., "survival-of-the-fittest", techniques. One of the oldest and well-known evolutionary algorithms are Genetic Algorithms (GA) (Slowik and Kwasnicka 2020). GA are widely applied as optimization algorithm in the field of model calibration (Park and Qi 2005; Malleson 2014). Classic GA are based on Darwin's theory of natural selection. The idea is that fittest individuals have a higher change to survive, and thus their genes contribute more to the reproduction of the next generation (Whitley 1994). Each parameter of the optimization represents a gene. Each solution of the optimization corresponds to a combination of genes, also known as a chromosome of an individual.

GA follow four steps: (1) initialization, (2) selection, (3) recombination, and (4) mutation (Mirjalili 2019). At the initialization, a random population to ensure diversity in the solution space is spawned. Next, a selection of the best solutions based on their fitness value is created. The fitness value is calculated by the user defined fitness function, i.e., the objective function of the optimization. After this, the chromosomes

are combined to produce new chromosomes, also called recombination. This means that two solutions (parents solutions) are selected to produce new solutions (children solutions). Cross-over operators are used to combine and swap the genes of the parent solutions to produce children solutions. In the last step, the genes of some children solutions are altered, also called mutation. In this way, the algorithm maintains the diversity of the population since a certain level of randomness is included in the population. This avoids the probability that GA stay in the local optimum (Mirjalili 2019).

GA are an iterative process, meaning that it keeps on creating new populations using selection, recombination, and mutation until some user defined stopping criterion is reached. In this research, we use the number of function evaluations as a stopping criterion.

2.2.3 Approximate Bayesian Computation

Model calibration is a core application of Bayesian data analysis using Bayes' theorem (Csilléry et al. 2010). In the case of sparse data and uncertainties, approximate Bayesian computing (ABC) is one of most suitable techniques for calibrating as it is likelihood-free (Vrugt and Beven 2018). ABC is a technique for estimating the posterior distribution of model parameters using Bayesian statistics.

One of the most efficient sampling algorithms for ABC is Differential Evolution Adaptive Metropolis (DREAM), a multi-chain Markov Chain Monte Carlo Sampling algorithm (Sadegh and Vrugt 2014). DREAM combines a multi-chain Markov Chain with differential evolution, as also found in some GA, for population evolution with a Metropolis selection rule. More specifically in the case of calibration, DREAM draws samples using the Markov Chain Monte Carlo Sampling method. These samples are used to run the simulation model, and to collect data. The distance between the simulated data and the observed data is used to either accept or reject a sample using an adaptive selection rule. DREAM uses multiple parallel chains to explore the solutions space adequately, and cross-over of solutions between the chains exists (Vrugt 2016).

The above steps in each chain are repeated until a stopping criterion, i.e., the number of draws, is reached. When this happens, the accepted samples are used to approximate the posterior parameter distribution.

2.3 Distance Metric

In order to minimize the difference between the simulation model data and the observed data, a so-called distance metric needs to be defined. The distance metric represent the distance between the simulation model data and the observed data given a certain function. Generally, standard statistical functions such as the mean square error, Kolmogorov-Smirnov metrics, or Euclidean (L2) distance are used as distance metric. However, most of these standard statistical functions do not properly adapt to the data of a specific problem (Suárez et al. 2021). In our case, the metric has to incorporate data of stochastic models in combination with sparse observed data of the system. Moreover, in simulating complex and large systems we typically deal with a high-dimensional space as a result of the many components with their parameters, which makes it challenging to find an appropriate and meaningful distance metric (Aggarwal et al. 2001).

Mirkes et al. (2020) note that the classic distance metric, such as L1 and L2, are highly efficient for complex and high-dimensional data applications. Thus, for the purpose of this study, we use a classic distance metric: the Manhattan (L1) distance. The Manhattan distance is the distance between data points as the sum of the absolute differences normalized for all dimensions.

3 DESIGN OF EXPERIMENTS USING GROUND TRUTH

We perform experiments to analyze the quality-of-fit of the selected model calibration techniques for different degrees of data sparseness. First, we explain the set-up for evaluating the quality-of-fit for the three selected model calibration techniques by using the ground truth. Next, we discuss the configuration of each technique.

3.1 Ground Truth Set-up for Evaluating the Quality-of-Fit

This research uses a ground truth set-up to evaluate the quality-of-fit of the model calibration techniques over various degrees of data sparseness. For replicating the observed data of the system, we use a simulation model that serves as a ground truth and extract data from this model. By using this set-up, we can assess how close the estimation of the calibration is to the true values as these are known. This is nearly impossible with real data (Khondoker et al. 2016).

Figure 1 presents the method used for evaluating the model calibration techniques. First, we define a ground truth simulation model with as input decision variable X with ground truth $X = x$. The output of the ground truth simulation model is the *ground truth data*, which does not include any sparseness. Next, we add data sparseness to the ground truth data. For example, 10% of the ground truth data elements are transformed into missing values. This leads to *sparse observed data*. The simulation model is calibrated to the *sparse observed data*. For the calibration, each iterative model calibration technique in essence selects a candidate value for the decision variable, $X = v$ (Frank et al. 2013). Five replications of the simulation model are ran based on the candidate values, leading to the *simulation model data* as output. Then, the distance between the *simulated model data* and the *sparse observed data* is calculated using the distance metric. This distance is minimized by the model calibration technique. Based on the distance, the model calibration technique selects new candidate values for the decision variable. This process stops when a stopping criterion is reached. The result is a value for the decision variable, $X = v^*$, that best describes the ground truth model, according to the model calibration technique.

Although the model calibration techniques minimizes the distance between the simulated and observed *output* data, the decision variable of the calibrated simulation model is not necessarily close to the decision variable of the ground truth model. So, we introduce the quality-of-fit of the decision variables which is defined as the normalized distance between the ground truth decision variable, $X = x$, and the optimal decision variable resulting from the simulation model calibration, $X = v^*$. This quality-of-fit is calculated by normalizing the difference between the ground truth input, $X = x$, and the solution, $X = v^*$, given the upper and lower bounds of the decision variable X. A quality-of-fit of 0 means that the optimal solution resulting from the calibration is not close to the ground truth; a quality-of-fit of 1 means that the optimal solution resulting from the calibration is the same as the ground truth.

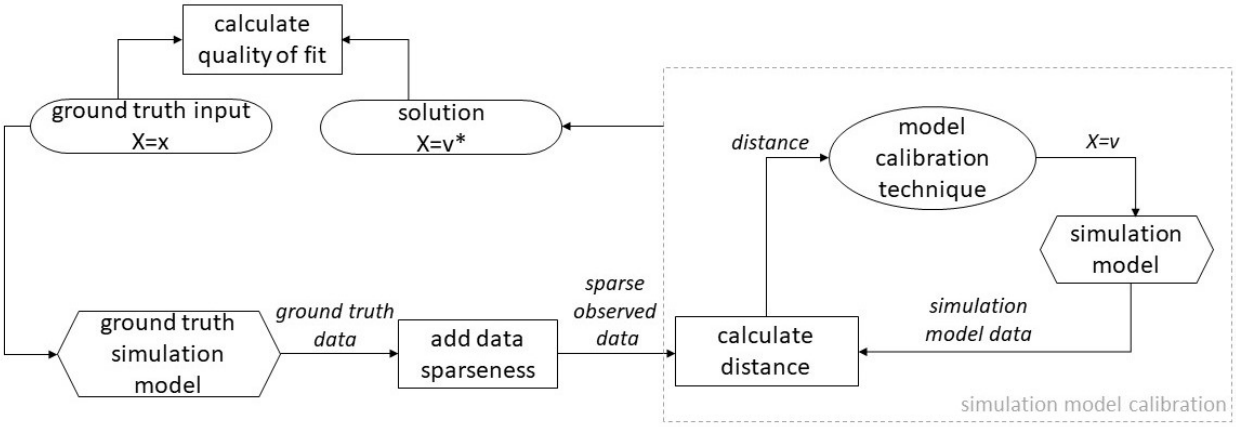


Figure 1: Method for evaluating calibration of a simulation model with sparse data.

The above steps represent one experiment for evaluating the quality-of-fit of a model calibration technique, given a certain degree of data sparseness. We systematically increase the degree of data sparseness added to the *ground truth data*. We evaluate for 10%, 25%, 50%, 75%, and 90% data sparseness. A degree of x% means that x% of the original data elements are missing values. It is randomly determined which x% of data elements are missing over the entire data set. Additionally, the model calibration techniques are examined

for 0% of data sparseness, i.e., *ground truth data*, as a base case. Each experiment is performed with 8 seeds to account for the effect of stochasticity on the quality-of-fit. For each seed, we first transform $x\%$ of the data set to missing values, and then we use this as input for all three model calibration techniques. This means that the exact same observations were left out of the data set that is presented to the different techniques for simulation model calibration.

3.2 Configuration of Model Calibration Techniques

To calculate the quality-of-fit for the selected model calibration techniques, the ground truth decision variable, $X = x$, is compared to the optimal solution, $X = v^*$, for each of the model calibration techniques. The result of Powell's Method and GA is a single optimal solution of the decision variable, so $X = v^*$. However, the result of ABC is an approximate posterior distribution of the decision variable. To extract one optimal value of decision variable X from this resulting posterior distribution, we select the value with the highest frequency, i.e., the mode, for that specific distribution. In this way, the most often accepted value of the decision variable represents the optimal solution for ABC as $X = v^*$.

For pragmatic reasons, we define a stopping criterion for finding the optimal solution for each technique. The stopping criteria for these experiments are based on an empirical analysis on the convergence of the model calibration techniques over 5 seeds. For the reference technique, Powell's Method, we limit the number of function evaluations to 1500 and the number of iterations to 100. For GA, we use 15.000 function evaluations as a stopping criterion. The analysis shows that with 15.000 function evaluations, the number of improvements stays constant for every seed. For ABC, we use 20.000 draws as the stopping criterion. The analysis shows that there is convergence of ABC determined by the Gelman-Rubin statistics at 20.000 draws for 3 of the 5 seeds (Gelman and Rubin 1992).

4 CASE STUDY: COUNTERFEIT PPE SUPPLY CHAIN

To evaluate the model calibration techniques, we use a case study of a counterfeit PPE supply chain. First, we introduce the stylized simulation model based on this case study. Next, we discuss the analysis and comparison of the various model calibration techniques given the simulation model of this case study.

4.1 Introduction of the Simulation Model

A discrete event simulation model of a stylized configuration of a counterfeit PPE supply chain from Vietnam to stores in the Netherlands is used. We assume that the counterfeit PPE are produced in Vietnam; one of the countries where most PPE come from, next to China and India. Most of these products are transported over sea to Europe following the legitimate transport flows. After arrival in Europe, they are distributed over various stores.

Figure 2 visualizes the stylized counterfeit PPE supply chain in more detail. The symbols represent the main actors in the supply chain, and the arrows represent the transportation flows. Starting from the supplier, supplies for PPE such as fabrics are delivered to the manufacturer over land in the production country, Vietnam. The manufacturer produces the counterfeit PPE in the factory and packs them in batches for transport. Each batch has a certain quantity of counterfeit PPE. For example, a batch consists of 1000 boxes of 100 PPE that equals a quantity of 100,000 PPE in total. Next, a truck transports a batch of finished counterfeit PPE to the export port in Hai Phong, Vietnam. The batch is loaded into a container and transported by a feeder to the transit port, Tanjung Pelepas, Malaysia. Once the batch is loaded on the feeder, it becomes part of the legitimate transport flow. At the transit port, the feeder unloads the container with counterfeit PPE. At the same port, the container is loaded onto a vessel, i.e., a larger container ship, for international transport. After a certain amount of days on international waters, the vessel arrives at the import port in Rotterdam, The Netherlands. The container is unloaded here, and waits for inland transport to the wholesales distributor in Eindhoven, The Netherlands. The wholesales distributor can also be seen as the stash location for the counterfeit PPE. At the wholesales distributor, the batch of counterfeit PPE

in the container is equally divided into three smaller batches for the retailers. These smaller batches are transported by small trucks to the retailer. When the counterfeit PPE arrive at the retailer, customers (either businesses or individual customers) can purchase the products with or without being informed that they are counterfeit.

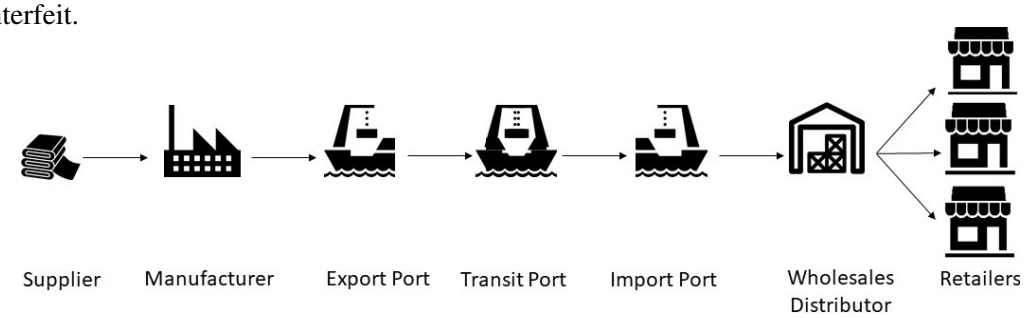


Figure 2: Visualization of the stylized counterfeit PPE supply chain.

The structure of the supply chain is linear. Due to the many uncertainties in the supply chain (e.g., delay in transport modalities, loading and unloading times), the supply chain becomes complex. For example, the retailer's inventory can fluctuate very much, depending on whether a vessel has a 1-day delay or a 7-day delay. In the simulation model, most uncertainties such as delays of transport modalities and speed of transport modalities follow triangular distributions inspired by real world data of a fashion retailer (Kuipers 2021).

In this research, the manufacturing duration, also referred to as manufacturing time, is the system parameter to be calibrated. More specifically, we use the manufacturing time as the decision variable in the simulation model calibration, meaning that we seek for the most likely value for the system parameter of manufacturing time. Table 1 shows the configuration of the manufacturing time as a decision variable. Manufacturing time has been chosen as an uncertain system parameter in this study for three reasons: (1) manufacturing time in another country is typically unobservable from the client's location, (2) there were many orders due to COVID-19 that could lead to extreme delays, and (3) delays at the beginning of the supply chain often have an unpredictably high impact on the rest of the supply chain due to the snowball effect.

Table 1: Configuration of manufacturing time.

Decision Variable	**Ground Truth**	**Lower Bound**	**Upper Bound**	**Unit**
Manufacturing Time (X)	2.5	1	10	Days

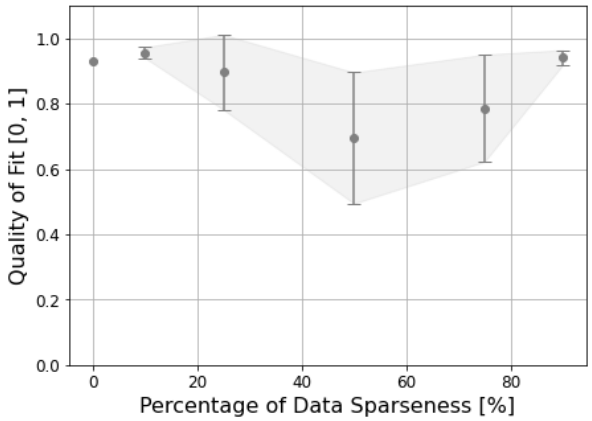
Given the value of the decision variable, in this case the manufacturing time, the simulation model is evaluated using a time series of the inventory levels of PPE for each actor in the supply chain (e.g., manufacturer, export port, import port) per day. The time series over multiple replications are combined using the mean values per day. Aggregated statistics of these combined time series are created, serving as the *simulation model data*. The statistics to represent the time series of each actor are the mean, standard deviation, 5th percentile, 95th percentile, and the average interval time (i.e., interval between the arrival of batches at actors). The data used for calibration includes the aggregated statistics of all actors.

The discrete event simulation model is developed with the library pydsol in Python. This library is a Python implementation of the Distributed Simulation Object Library (DSOL), originally implemented in Java (Jacobs 2005).

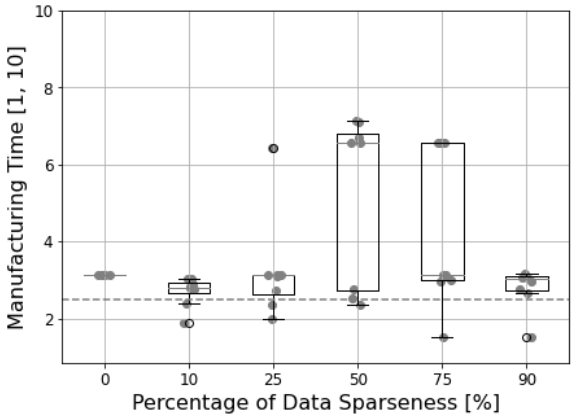
4.2 Analysis of Powell's Method, GA & ABC

We analyze the quality-of-fit for the reference technique, Powell's Method, and the selected techniques, GA and ABC, given certain degrees of data sparseness and using 8 replications with unique seeds. For each technique, we show a graph of the average quality-of-fit with a 95% confidence interval to visualize

the spread of the solutions over various replications. Besides, we show a boxplot of the calculated optimal values of the decision variable *manufacturing time* resulting from the various replications. In addition, a table is presented to compare the reference technique and the selected model calibration techniques by the average quality-of-fit and the standard deviation.



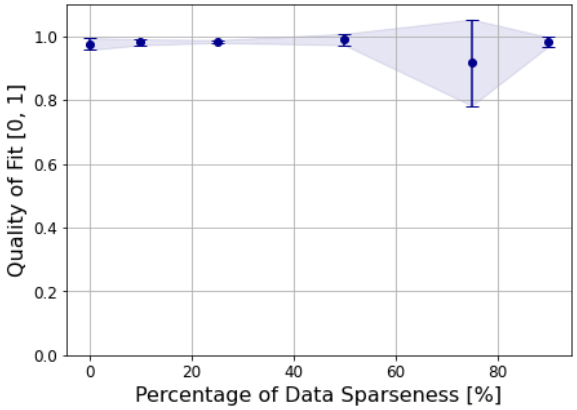
(a) Average quality-of-fit with the 95% confidence interval.



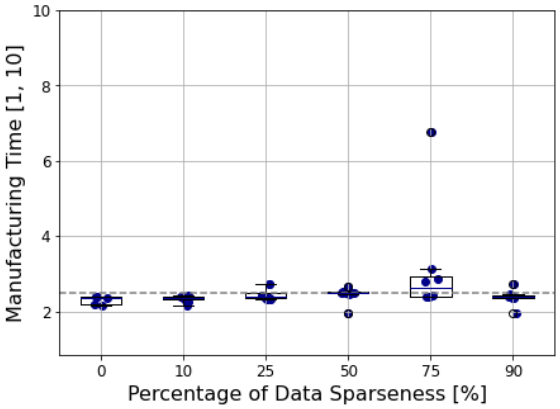
(b) Boxplot of the optimal values of manufacturing time. The dashed gray line is the ground truth value of manufacturing time: 2.5 days.

Figure 3: Results for Powell's Method for 8 seeds for various degrees of data sparseness.

Figure 3a shows that Powell's Method has an average quality-of-fit between 0.70 to 0.96. When data sparseness is more than 10%, the average quality-of-fit decreases and the 95% confidence interval becomes wider. Figure 3b shows that from 10% data sparseness onward, the algorithm finds optimal values of more than 6 days for the manufacturing time. Interestingly, there are no optimal values found between 3 and 6 days. Surprisingly, Powell's method has a high quality-of-fit with a small confidence interval with 90% data sparseness.



(a) Average quality-of-fit with the 95% confidence interval.

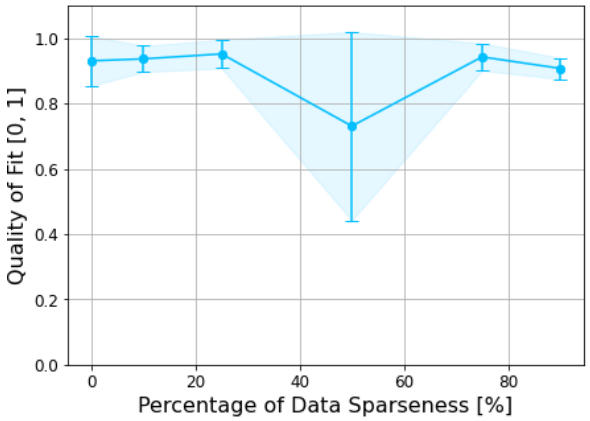


(b) Boxplot of the optimal values of manufacturing time. The dashed gray line is the ground truth value of manufacturing time: 2.5 days.

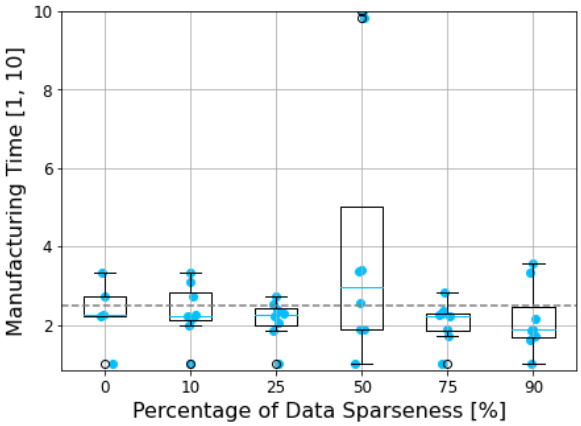
Figure 4: Results for Genetic Algorithm for 8 seeds for various degrees of data sparseness.

Figure 4a shows that GA has an average quality-of-fit between 0.92 and 0.99. The quality-of-fit and the correlated spread stays constant for most of the chosen values for data sparseness. The 95% confidence interval is narrow for the different degrees of data sparseness. Only with 75% data sparseness, there are

more solutions that have a lower quality-of-fit and the 95% confidence interval is wider. In Figure 4b, we see that for 75% data sparseness, most optimal solutions for the decision variable are slightly above the ground truth value. There is one outlier where the optimal manufacturing time is calculated to be more than 6 days.



(a) Average quality-of-fit with the 95% confidence interval.



(b) Boxplot of the optimal values of manufacturing time. The dashed gray line is the ground truth value of manufacturing time: 2.5 days.

Figure 5: Results for approximate Bayesian computation for 8 seeds for various degrees of data sparseness.

Figure 5a shows that ABC has an average quality-of-fit between 0.78 and 0.98. For most of the degrees of data sparseness, the average quality-of-fit is around 0.95 and the 95% confidence interval is narrow. Only at 50% data sparseness, the average quality-of-fit is the lowest, i.e., around 0.78, and the confidence interval is relatively wide. Figure 5b shows that at 50% data sparseness, the algorithm has a wide spread of optimal solutions for the value of manufacturing time. Some solutions are close to the lower bound and the upper bound of this decision variable. Other solutions are closer to the ground truth value, i.e., between 3 and 4 days, but are still relatively far from the ground truth compared to experiments with other degrees of data sparseness. The resulting posterior distribution of the algorithm for 50% data sparseness follows a bimodal distribution.

Table 2: Quality-of-fit in mean and standard deviation for each model calibration techniques for various degrees of data sparseness.

	Powell's Method		**Genetic Algorithm**		**Approximate Bayesian Computation**	
Percentage of Data Sparseness	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*
0%	0.93	0.00	0.98	0.01	0.93	0.05
10%	0.96	0.02	0.98	0.01	0.94	0.04
25%	0.90	0.13	0.98	0.00	0.95	0.04
50%	0.70	0.22	0.99	0.02	0.73	0.32
75%	0.79	0.18	0.92	0.15	0.94	0.05
90%	0.94	0.03	0.98	0.02	0.91	0.04

Table 2 presents the results of the average quality-of-fit and the corresponding standard deviation of the reference technique and the two selected model calibration techniques for various degrees of data sparseness. It shows that GA outperforms Powell's Method and ABC for all percentages of data sparseness in terms of a higher average quality-of-fit and a lower standard deviation. ABC performs slightly better on the average quality-of-fit than Powell's Method. However, Powell's Method and ABC both have a relatively high standard deviation compared to GA, meaning that there is more variation in the distance of the optimal solution to the ground truth value. Over the various degrees of data sparseness, Powell's Method has the highest standard deviation. It is quite remarkable that Powell's Method and ABC have the lowest average quality-of-fit and the highest standard deviation for 50% data sparseness. For both techniques, the average quality-of-fit increases again for 75% and 90% data sparseness.

Overall, GA and ABC outperform the reference technique for calibrating the counterfeit PPE supply chain simulation model over various values for data sparseness. From this analysis, GA shows to be the most promising for calibrating a simulation model with sparse data due to the high average quality-of-fit, the narrow 95% confidence interval, and a small standard deviation over all degrees of data sparseness.

5 DISCUSSION

Overall, the results show that the selected model calibration techniques seem to have a high quality-of-fit for calibrating the counterfeit PPE simulation model with sparse data. There are three limitations for generalizing the results: (1) local vs. global optimum, (2) specific to supply chains, and (3) lack of including structural uncertainty and of including other dimensions of data sparseness.

Regarding the local vs. global optimum, it is remarkable that Powell's Method and ABC both have the lowest quality-of-fit and the highest standard deviation at 50% data sparseness. A possible explanation for this result for Powell's Method is that the algorithm sometimes gets stuck in a local optimum, instead of reaching the global optimum (Powell 1964), possibly caused by two input spaces of interest. A possible explanation for ABC is that the algorithm results in a bimodal distribution, with more than one region in the input space that results in optimal solutions. Calibration with these algorithms can yield multiple counterfeit PPE supply chains that could represent the real world supply chain to a certain extent. We should therefore be careful in choosing the configuration to gain insights from. Not doing so could lead to a "wrong" view on criminal activities in the real world counterfeit PPE supply chain. In addition, a wider set of optimization algorithms could be explored for their effectiveness in model calibration.

The second limitation is that the results could be specific to the linear counterfeit PPE supply chain model. In general, a supply chain is often presented as a sequential network. This means, for example, that there is an one-directional flow between the supplier and the manufacturer. On the one hand, this direct and linear dependency between the actors could lead to more straightforward calibration of the simulation model with sparse data. This challenges the generalizability of the results to other systems. The linear supply chain also has a single parameter that needed to be calibrated, where in real situations, data of multiple parameters might be sparse. On the other hand, the results of this paper give a proof of concept on how data sparseness effects the ability to calibrate a linear supply chain using sparse data.

The third limitation is that the lack of including structural uncertainty and of including other dimensions of data sparseness. Keeping the structure of the simulation model the same for the ground truth and the calibrated model could be a crucial element for being able to find the optimal value for the parameter(s). When structure is included as a parameter, this could mean that it is more difficult for the model calibration techniques to converge to a solution with a high degree of data sparseness. In our example case, data sparseness in the form of missing data values were random, where in reality there could be patterns, such as missing data only during the night. Finally, data sparseness consists of more dimensions than missing values: examples are noise and bias. The effect of these other types of data sparseness on calibration quality is still unknown, making it difficult to generalize the results to all types of data sets. Nonetheless, this study gives insight in the quality-of-fit for one parameter when increasing the percentage of missing values, a type of uncertainty that often occurs in criminal cases, specifically during COVID-19.

6 CONCLUSION AND FUTURE WORK

This research is a first attempt to analyze the quality-of-fit of model calibration techniques that are likely to be suitable for calibrating simulation models in the case of sparse data. Due to the high data sparseness in counterfeit PPE supply chains, we used a PPE supply chain as our case study. We selected a reference technique that is often used for calibration of simulation models: Powell's Method. We selected GA and ABC as model calibration techniques that are likely to be suitable in case of sparse data. By using a ground truth set-up for evaluating the quality-of-fit, we assessed how accurately the three model calibration techniques find the optimal system parameter value for the simulation model with an increasing degree of data sparseness. The results demonstrate that the selected model calibration techniques are suitable for calibrating simulation models when faced with sparse data, at least for a linear supply chain with randomly missing values. This shows that with sparse data due to COVID-19 and criminals masking their data, the selected model calibration techniques can help to gain insight in underlying counterfeit PPE supply chains.

The main directions for future research are including more model calibration techniques, evaluating for a larger set of simulation models, introducing structural uncertainty and other dimensions of data sparseness.

REFERENCES

Aggarwal, C. C., A. Hinneburg, and D. A. Keim. 2001. "On the Surprising Behavior of Distance Metrics in High Dimensional Space". In *International Conference on Database Theory*, edited by J. Van den Bussche and V. Vianu, 420–434. London, UK: Springer.

Csilléry, K., M. G. Blum, O. E. Gaggiotti, and O. François. 2010. "Approximate Bayesian Computation (ABC) in Practice". *Trends in Ecology & Evolution* 25(7):410–418.

de Groot, L., and A. Hübl. 2021. "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 *Proceedings of the 2021 Winter Simulation Conference*, edited by S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, 1–12. Phoenix, Arizona: Institute of Electrical and Electronics Engineers, Inc.

De Santis, A., T. Giovannelli, S. Lucidi, M. Messedaglia, and M. Roma. 2022. "A Simulation-Based Optimization Approach for the Calibration of a Discrete Event Simulation Model of an Emergency Department". *Annals of Operations Research*:1–30.

Frank, M., C. Laroque, and T. Uhlig. 2013. "Reducing Computation Time in Simulation-Based Optimization of Manufacturing Systems". In *Proceedings of the 2013 Winter Simulations Conference*, 2710–2721. Washington, District of Columbia: Institute of Electrical and Electronics Engineers, Inc.

Gelman, A., and D. B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences". *Statistical Science* 7(4):457–472.

Hazen, B. T., C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer. 2014. "Data Quality for Data Science, Predictive Analytics, and Big Data in Supply Chain Management: An Introduction to the Problem and Suggestions for Research and Applications". *International Journal of Production Economics* 154:72–80.

Hofmann, M. 2005. "On the Complexity of Parameter Calibration in Simulation Models". *The Journal of Defense Modeling and Simulation* 2(4):217–226.

Huang, Y. 2013. *Automated Simulation Model Generation*. Ph.D. Thesis, Faculty of Technology, Policy and Management, Delft University of Technology, Delft, Netherlands.

Jacobs, P. H. M. 2005. *The DSOL Simulation Suite*. Ph.D. Thesis, Faculty of Technology, Policy and Management, Delft University of Technology, Delft, Netherlands. http://resolver.tudelft.nl/uuid:4c5586e2-85a8-4e02-9b50-7c6311ed1278.

Jalali, H., I. Van Nieuwenhuyse, and V. Picheny. 2017. "Comparison of Kriging-Based Algorithms for Simulation Optimization with Heterogeneous Noise". *European Journal of Operational Research* 261(1):279–301.

Khondoker, M., R. Dobson, C. Skirrow, A. Simmons, and D. Stahl. 2016. "A Comparison of Machine Learning Methods for Classification Using Simulation with Multiple Real Data Examples from Mental Health Studies". *Statistical Methods in Medical Research* 25(5):1804–1823.

Kuipers, L. 2021. "Increasing Supply Chain Visibility With Limited Data Availability: Data Assimilation In Discrete Event Simulation". Master's thesis, Faculty of Technology, Policy and Management, Delft University of Technology, Delft, Netherlands. https://resolver.tudelft.nl/uuid:5f68b82f-205e-4509-9a64-22082c46065f.

Liu, Z., D. Rexachs, F. Epelde, and E. Luque. 2017. "A Simulation and Optimization Based Method for Calibrating Agent-Based Emergency Department Models Under Data Scarcity". *Computers & Industrial Engineering* 103:300–309.

Malleson, N. 2014. "Calibration of Simulation Models". *Encyclopedia of Criminology & Criminal Justice* 40:115–118.

Mirjalili, S. 2019. "Genetic Algorithm". In *Evolutionary Algorithms and Neural Networks. Studies in Computational Intelligence*, Volume 780, 43–55. Cham: Springer.

Mirkes, E. M., J. Allohibi, and A. Gorban. 2020. "Fractional Norms and Quasinorms Do Not Help to Overcome the Curse of Dimensionality". *Entropy* 22(10):1–31.

Omar, I. A., M. Debe, R. Jayaraman, K. Salah, M. Omar, and J. Arshad. 2022. "Blockchain-Based Supply Chain Traceability for COVID-19 Personal Protective Equipment". *Computers & Industrial Engineering* 167:107995.

Ören, T. I. 1981. "Concepts and Criteria to Assess Acceptability of Simulation Studies: A Frame of Reference". *Communications of the Association for Computing Machinery* 24(4):180–189.

Park, B., and H. Qi. 2005. "Development and Evaluation of a Procedure for the Calibration of Simulation Models". *Transportation Research Record* 1934(1):208–217.

Powell, M. J. 1964. "An Efficient Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives". *The Computer Journal* 7(2):155–162.

Puchinger, J., and G. R. Raidl. 2005. "Combining Metaheuristics and Exact Algorithms in Combinatorial Optimization: A survey and Classification". In *First International Work-Conference on the Interplay Between Natural and Artificial Computation*, 41–53. Las Palmas, Spain: Springer.

Ren, Y., and Y. Wu. 2013. "An Efficient Algorithm for High-Dimensional Function Optimization". *Soft Computing* 17(6):995–1004.

Sadegh, M., and J. A. Vrugt. 2014. "Approximate Bayesian Computation Using Markov Chain Monte Carlo Simulation: DREAM (ABC)". *Water Resources Research* 50(8):6767–6787.

Schmitt, A., and M. Singh. 2009. "Quantifying Supply Chain Disruption Risk Using Monte Carlo and Discrete-Event Simulation". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. Rossetti, R. R. Hill, and B. Johansson, 1237 – 1248. Austin, Texas: Institute of Electrical and Electronics Engineers, Inc.

Shannon, R. E. 1998. "Introduction to the Art and Science of Simulation". In *Proceedings of the 1998 Winter Simulation Conference*, edited by D. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, 7–14. Washington, District of Columbia: Institute of Electrical and Electronics Engineers, Inc.

Slowik, A., and H. Kwasnicka. 2020. "Evolutionary Algorithms and Their Applications to Engineering Problems". *Neural Computing and Applications* 32(16):12363–12379.

Suárez, J. L., S. García, and F. Herrera. 2021. "A Tutorial on Distance Metric Learning: Mathematical Foundations, Algorithms, Experimental Analysis, Prospects and Challenges". *Neurocomputing* 425:300–322.

Vassiliadis, V. S., and R. Conejeros. 2009. "Powell Method". In *Encyclopedia of Optimization*, edited by C. A. Floudas and P. M. Pardalos, 3012–3013. Boston, MA: Springer.

Vrugt, J. A. 2016. "Markov Chain Monte Carlo Simulation Using the DREAM Software Package: Theory, Concepts, and MATLAB Implementation". *Environmental Modelling & Software* 75:273–316.

Vrugt, J. A., and K. J. Beven. 2018. "Embracing Equifinality with Efficiency: Limits of Acceptability Sampling Using the DREAM (LOA) Algorithm". *Journal of Hydrology* 559:954–971.

Whitley, D. 1994. "A Genetic Algorithm Tutorial". *Statistics and Computing* 4(2):65–85.

Wigan, M. R. 1972. "The Fitting, Calibration, and Validation of Simulation Models". *Simulation* 18(5):188–192.

Xie, X. 2018. *Data Assimilation in Discrete Event Simulations*. Ph.D. Thesis, Faculty of Technology, Policy and Management, Delft University of Technology, Delft, Netherlands.

Zhong, J., and W. Cai. 2015. "Differential Evolution with Sensitivity Analysis and the Powell's Method for Crowd Model Calibration". *Journal of Computational Science* 9:26–32.

AUTHOR BIOGRAPHIES

ISABELLE M. VAN SCHILT is a Ph.D. candidate at the Policy Analysis section of the Faculty of Technology, Policy and Management at Delft University of Technology. Her research focuses on the use of simulation models for calibrating complex systems that are characterized by uncertainty and sparse data (i.e., counterfeit supply chains). Her email address is i.m.vanschilt@tudelft.nl.

JAN H. KWAKKEL is a full professor of decision-making under deep uncertainty at the Policy Analysis section of the Faculty of Technology, Policy and Management at Delft University of Technology. His research interests are in the field of the developing and testing innovative model-based techniques in deep uncertainty situations. His email address is j.h.kwakkel@tudelft.nl.

JELTE P. MENSE is a senior researcher at the National Policelab AI at the Dutch Police and visiting researcher at the Department of Information and Computing Sciences at Utrecht University. His current work focuses on the application of artificial intelligence in law enforcement. His email address is j.p.mense@uu.nl.

ALEXANDER VERBRAECK is a full professor of systems and simulation at the Policy Analysis section of the Faculty of Technology, Policy and Management at Delft University of Technology. Also, he has a position as adjunct professor at the R.H. Smith School of Business at the University of Maryland, USA. His research focuses on complex systems and simulation, especially in the field of discrete event simulation and logistics. His email address is a.verbraeck@tudelft.nl.