Simulation-based inference in agent-based models using spatio-temporal summary statistics

Eric Dignum[1 [0000−0002−9560−8005]], Harshita Choudhary[1], and Mike Lees[1]

University of Amsterdam, Science Park 900, 1098 XH Amsterdam
e.p.n.dignum@uva.nl

Abstract. In agent-based models (ABMs), traditional statistical inference faces challenges due to intractable likelihoods and computational costs. This study evaluates neural posterior estimation (NPE) and neural ratio estimation (NRE) for parameter inference in ABMs and compares them with approximate Bayesian computation (ABC). NPE and NRE are argued to be more efficient than traditional methods such as ABC and circumvent some of their limitations. The assessment of the methods focuses on the satisfaction threshold in Schelling's model of residential segregation, including regions of high variance and non-equilibrium dynamics. As these simulation-based methods still require summary statistics as high-level descriptions of the ABM, we propose a general approach to construct them based on spatial and/or temporal information and evaluate how the different summary statistics affect performance. Both NPE and NRE generally outperform ABC regardless of summary statistics. Most notably, NRE excels when employing the most detailed spatio-temporal information, but adding spatial or temporal information alone is not always beneficial for NPE, NRE and ABC. This holds true for different training budgets and when estimating multiple parameters. Hence, the study underscores the importance of spatio-temporal information for accurate parameter inference in this ABM, but information redundancy can degrade performance as well. Therefore, finding optimal high-level descriptions to capture fundamental emergent patterns in the model through summary statistics might prove crucial in cases where the systems are governed by more complex behaviour.

Keywords: simulation-based inference · agent-based modelling · neural posterior estimation · neural ratio estimation · approximated bayesian computation

1 Introduction

Traditional statistical methods for parameter inference heavily rely on likelihood functions. However, in many cases, models reach a level of complexity where the likelihood becomes impossible to derive exactly or even sample from. In such cases, simulation-based inference (SBI) methods have emerged as a powerful tool for parameter inference when it is still possible to generate data from the model under study. Recently, there has been an intersection of deep learning with

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

the field of SBI, resulting in novel neural network-based inference approaches [2]. One modelling paradigm that might benefit from these novel methods is agent-based modelling. Agent-based models (ABMs) often consist of many (heterogeneous) agents with rule-based actions that interact with each other and their environment. This can result in macro-scale emergent outcomes that are often not expected when studying agents in isolation [3]. Using mathematical modelling and computational algorithms, one can simulate the behaviour of agents, their interactions, and consequential actions, all of which influence the overall dynamics of the system [1]. While a flexible modelling paradigm, due to interdependent and complicated behavioural rules, non-linear dynamics, and inherent stochasticity, it is often very difficult or impossible to perform likelihood-based inference on. As these new SBI techniques can identify complex relationships efficiently, they seem to be very suitable for calibrating ABMs on data [5].

Approximate Bayesian computation (ABC) [16] is a SBI method that tries to approximate the posterior, the values of the parameters that are most likely to have generated the observed data, based on the discrepancy between the simulated (model-generated) data and the observed data. It is common practice in SBI to use summary statistics which provide a condensed representation of model behaviour, capturing essential information while reducing the dimensionality. ABC in particular requires choosing a distance metric and a threshold; a parameter value is accepted to come from the posterior distribution if the discrepancy between simulated and observed summary statistics falls within the threshold or filter those with the smallest distances to the observed data. However, deciding the distance metric and the threshold value can greatly affect performance [5]. Moreover, ABC aims to estimate an approximate posterior, the accuracy of which depends on the chosen threshold, and inference is only valid for the specific observation used; hence, the procedure has to be repeated for different (empirical) observations of the same system. The newly proposed neural network-based approaches aim to learn the relationship between the parameter values and their corresponding summary statistics. This addresses some of the limitations encountered in traditional ABC techniques. They eliminate the need to choose a distance metric and threshold, and they are amortised, meaning they can estimate posterior probabilities for any new observation not seen by the network. Lastly, they do not throw away samples as in ABC, but they use all available data. This can lead to a more efficient use of simulations, which is especially important when these are computationally expensive, which is typically the case for ABMs. However, this new generation of inference techniques has yet to be extensively tested for parameter inference in ABMs, nor has it been simultaneously compared with ABC. In addition, summary statistics still have to be selected carefully for the calibration methods, and a general approach to do this for ABMs is currently missing.

In this paper, we propose a general approach to construct summary statistics for spatio-temporal ABMs, while comparing the performance of neural posterior estimation (NPE) [6,12] and neural ratio estimation (NRE) [7,11] with ABC. NPE and NRE have shown promising results for economic ABMs, producing sig-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

nificantly more accurate parameter estimations while requiring fewer simulations [5], but only using a specific configuration of the models tested. We demonstrate our approach by calibrating the entire range of the satisfaction threshold parameter of Schelling's model of residential segregation [13]. This model is chosen as a test case as it has only one parameter that governs segregation dynamics and its underlying mechanisms are well described. Depending on the value of the threshold, one can observe emergent behaviour, non-linear patterns, non-equilibrium dynamics and/or regions with a large variance in outcomes. These can provide challenging cases for the calibration methods while still having relatively easy-to-understand dynamics. Our approach is general enough that it can be applied to other (spatial) ABMs, and our results provide some indication of how the inference methods may perform in other scenarios. Both NPE and NRE generally outperform ABC regardless of summary statistics. Most notably, NRE excels when employing the most detailed spatio-temporal information, but adding spatial or temporal information alone is not always beneficial for the methods. Hence, the study underscores the importance of spatio-temporal information for accurate parameter inference in this ABM, but information redundancy can degrade performance as well. Therefore, finding optimal high-level descriptions to capture fundamental emergent patterns in the model through summary statistics might prove crucial in cases where the systems are governed by more complex behaviour.

2 Background and related work

In this section, only three approaches for calibrating simulation models are discussed. For a detailed overview of existing methods, we refer to [2]. In SBI the aim is to estimate the posterior probability distribution $P(\theta|X)$, i.e., the posterior distribution of the model parameters (θ) conditional on both the simulated data from the model (X^{sim}) and potentially the observations of the real system (X^{obs}). When the likelihood $P(X|\theta)$ is intractable, but it is possible to simulate X from a generative model given a set of parameter values (θ), one can perform SBI. One of the most commonly used techniques in this area is approximate Bayesian computation (ABC).

2.1 Approximate Bayesian Computation

ABC has been applied to the calibration of ABMs in various fields, including economics [8], epidemiology [18], and cancer research [14]. In this section, only the most basic version of ABC will be described. However, for more details, [9] provide an elaborate study on the various improvements and extensions of ABC algorithms. The general idea of ABC methods is to generate samples from the posterior distribution by simulating data $X^{sim} \sim p(X|\theta)$ and to assess whether the simulated data are close to the observed data. Specifically, if the discrepancy between X^{sim} and X^{obs} according to a distance metric d falls within a certain acceptance threshold ϵ. Note that this makes posterior inference using ABC only

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

valid for the specific observation (X^{obs}) used, which is also called: non-amortised. Rejection ABC is the most basic algorithm of the ABC methods [16]. Here, θ_i is randomly sampled from a prior distribution $P(\theta)$. Those θ_i for which the distance between X^{sim} and X^{obs} is less than ϵ are accepted as samples from the posterior distribution. General ABC rejection for sampling one parameter from the posterior is as follows:

1. Sample $\theta_i \sim P(\theta)$
2. Run the generative model with θ_i as input and save summary statistics X_i^{sim} that are a description of the model behaviour
3. Accept θ_i as a sample coming from the posterior $P(\theta|X)$ if $d(X_i^{sim}, X^{obs}) \leq \epsilon$
4. Repeat this until a specific number of accepted samples or total simulations is reached

This technique draws independent samples from the approximate posterior. However, there are some drawbacks to this approach. First, one has to choose an appropriate distance metric and a value for ϵ. Second, the acceptance rate is typically low, specifically when the posterior is much narrower than the prior distribution, and third, it can be computationally very expensive for a small value of ϵ. In this study, 10% of the simulations with the smallest Euclidean distances to the observed data are kept to estimate the posterior distribution.

2.2 Neural density estimation

To avoid the dependence on ϵ, [12] proposed NPE, that frames parameter inference as a conditional density estimation problem. The method takes the input X and produces the posterior distribution $P(\theta|X)$ by training a neural network (NN) on simulated data X^{sim} and the corresponding parameter vector θ. The NN tries to learn the probabilistic relationship between X^{sim} and θ through this training. Note that the empirical data has not been utilised yet and the NN essentially acts as a surrogate for the generative model. This approach amortises the inference process, involves training a neural conditional density estimator once, using training data consisting of data-parameter pairs (X_i^{sim}, θ_i) where $X_i^{sim} \sim P(X|\theta = \theta_i)$, and then condition the posterior distribution for any X^{obs}. This is an improvement over ABC, as ABC's inference is only valid for the specific X^{obs} used. Various variants were developed, but we use the NPE-C variant of [6] as it is the most flexible and best performing approach.

NRE takes a different approach and uses supervised classification [11]. In its simplest form, it works as follows: create a random dataset of independent pairs by shuffling the data-parameter pairs as described above. A neural network is then trained to classify which combinations belong to the dependent dataset and which ones to the independent one. Specifically, NRE uses a neural network as a classifier to distinguish between dependent data points $(X_i^{sim}, \theta_i) \sim P(X, \theta)$ and independent data points $(X_i^{sim}, \theta_j) \sim P(X)P(\theta)$. The dependent pairs are generated from the simulator, while the independent pairs can be obtained by shuffling the (X_i^{sim}, θ_i) pairs. This destroys the dependencies, thus associating

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

X_i^{sim} with a random θ_j. Using the likelihood-to-evidence ratio, which represents the likelihood that the pair belongs to the dependent dataset, one can obtain the posterior distribution with the help of MCMC sampling. Similarly to NPE, NRE is an ϵ-free inference technique that does not require the acceptance-rejection step. It is also simulation-efficient (it does not reject simulations) and does not rely on a distance function. However, unlike NPE, direct sampling from the posterior is not feasible, as likelihood ratios are calculated. This necessitates the use of a sampling technique such as MCMC, introducing an additional, more computationally expensive step in the process. The variant of [11] is employed here (NRE-C) as it performs better than other variants in their experiments.

Both NPE and NRE have been used for parameter inference in ABM by [5] where they show similar or even improved performance compared to kernel density estimation techniques with fewer simulations in an economic ABM. However, they do not systematically sweep the parameter space. Hence, there might be parts where the techniques struggle to infer the parameters of the models. Moreover, they employ one set of self-constructed summary statistics and use a NN to learn a set of summary statistics from the simulations, but it is not clear how varying summary statistics more systematically affect the methods.

3 Methodology

This section provides a detailed explanation of our specific implementation of the Schelling model, as it is important to relate it to the performance of the calibration methods. Subsequently, the setup of several experiments with different summary statistics for the calibration methods is described, followed by the calibration pipelines for ABC, NPE and NRE to estimate the true value of the satisfaction threshold.

3.1 Schelling's ABM of residential segregation

To investigate the phenomenon of racial segregation, Schelling [13] developed one of the first ABMs. He showed that even a mild preference for having people of the same group in your local neighbourhood could result in a highly segregated society. Hence, the model yielded counter-intuitive findings, demonstrating that the outcome of the collective behaviour of agents could differ from the intentions of the isolated individuals, due to the non-linearity caused by the interacting agents. Although not completely similar to the original, below we describe the version used in this study.

- **Initialisation:** Agents are randomly placed on an 80×80 grid and their total number is such that the density is 90% to allow for relocation. With a probability 0.5, an agent belongs to either the blue (B) or orange (O) group.
- **Movement:** At each time step, 15% of agents are allowed to move. An agent is satisfied in their current location if the fraction of individuals of the same group in their 8 surrounding cells (Moore neighbourhood) is above a certain

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

threshold (μ_h). If an agent is not satisfied, they move to a randomly chosen vacant cell, otherwise they stay in their location. The tolerance level of each household (μ_h) is sampled from a truncated normal distribution with mean μ and standard deviation of 0.05.
- **Stopping:** The simulation is stopped after 100 time steps (to reduce computation time) or when every agent is satisfied.

The relocation process in the model occurs iteratively over time, where the movement of dissatisfied agents can subsequently affect the satisfaction levels of agents in their new neighbourhood. This potentially triggers a cascade of relocations. Segregation is measured as the average fraction of similar neighbours. For every household, the fraction of similar neighbours (the eight surrounding cells) is calculated, itself excluded, and averaged over all households. Although numerous metrics have been proposed to quantify segregation [10], the average fraction of similar neighbours is selected because of straightforward use and interpretation.

This simple but powerful model demonstrates that segregation can be much higher than would be needed to satisfy individual preferences. Figure 1 illustrates the results of the Schelling model with different tolerance levels, denoted by μ. These tolerance values indicate the threshold required for agent satisfaction. When $\mu = 0.3$, an agent is considered satisfied if at least 30% of its immediate neighbours, here considering a neighbourhood size of 8, belong to the same group. Note that satisfaction threshold, tolerance parameter, and similar terms are used interchangeably and refer to the parameter μ.

3.2 Experimental design

ABMs are time-driven models and often have an explicit spatial context [1]. We thus propose and base our experiments on a generic approach of taking summary statistics at different points (or scales) in time and space. With the Schelling model, a series of experiments is undertaken across the entire range of the tolerance parameter $\mu \in \{0.1, 0.2, ..., 0.8\}$. Both 0 and 0.9-1 are excluded as they are very similar to those of 0.1 and 0.8 respectively.

This assessment allows us to analyse the effectiveness of both methods in various regimes of model behaviour, including stable regions, non-equilibrium dynamics, and regions with substantial variance in observations to test the accuracy of ABC, NPE and NRE in different circumstances and compare them. In this context, the summary statistics are expected to describe the dynamics of the model sufficiently, such that one can infer the true values of the parameters that generated the data. Relatively low-dimensional summary statistics are often comparatively easy to interpret and can serve as a good indicator of the overall dynamics of the model. However, it might also lead to the loss of important information and affect the performance of the calibration methods. Therefore, several summary statistics are tested to see how this affects performance of ABC, NPE and NRE.

- **Scalar**: A scalar value that quantifies the level of segregation (average fraction of similar neighbours) observed at the end of one simulation ($t = 100$).

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

- **Temporal**: A 5-dimensional vector that represents the segregation evaluated in time steps $t = 20, t = 40, \ldots, t = 100$.
- **Spatial**: The average fraction of similar neighbours is calculated with a Moore neighbourhood of sizes 1, 2, 3, 4 and 5 at $t = 100$. The relocation rule for the agents is still based on a radius of 1.
- **Spatio-temporal**: Combines both spatial and temporal dimensions. A 25-dimensional vector that includes segregation observed at time steps (20, 40, ..., 100) and within various radii values (1, 2, 3, 4, and 5).

The training budget for the methods is varied ($N \in \{500, 1000, 2000\}$) to see how this affects the performance for the different summary statistics. The spatio-temporal summary statistics, containing 25 values, might benefit most from increasing the training budget. With lower sample sizes, inference methods could struggle to approximate posteriors in this high-dimensional space. As the implemented ABC-method retains 10% of the samples, the number of samples drawn from the posterior estimations of NPE and NRE are changed accordingly, to make it a fair comparison (i.e., $M \in \{50, 100, 200\}$). Moreover, a heterogeneous version of Schelling's model, where the two groups have different μ values, is tested. In this case, the methods have to infer two parameters and possibly a more complicated relationship between the tolerance parameters and the observed level of segregation. Note that we calculate the level of segregation at different spatio-temporal scales, but this can be any other metric that is of interest. For example, in an epidemiological model, the percentage of infections on the different scales could be used instead of segregation.

In summary, ABC, NPE and NRE will be used to estimate the satisfaction threshold (μ) of the Schelling model described in Section 3.1. The parameter represents the preference to have at least a certain fraction (μ) of similar neighbours. Its value ranges between 0.1 and 0.8, where 0.8 signifies a high degree of intolerance toward other groups, while 0.1 indicates impartiality or (almost) lack of preference. All three methods try to approximate $P(\mu|X)$ using the simulated data (X_i^{sim}, μ_i) generated from the model and an observation X^{obs}. In our case, X^{obs} is also generated by the model and setting $\mu = \mu_{true}$, but in real settings μ_{true} is generally not known. The SBI Python package is used to implement ABC, NPE and NRE with their default architectural specifications [15]. The inference procedure for ABC can be found in Section 2.1 and for NPE/NRE it can be summarised as follows:

1. Randomly sample $N \in \{500, 1000, 2000\}$ values from $P(\mu) \sim \mathcal{U}(0, 1)$.
2. Run the Schelling model with the input parameters μ_i for each $i \in \{1, 2, ..., N\}$ and save the summary statistics of each run.
3. Train the NN for NPE and NRE on the simulated data pairs (X_i^{sim}, μ_i).
4. Sample $M \in \{50, 100, 200\}$ times from the approximated posterior distribution and calculate the root mean squared error (RMSE) given μ_{true}: $RMSE = \frac{1}{M}\sqrt{\sum_{i=1}^{M}(\hat{\mu}_i - \mu_{true})^2}$.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

4 Results

Given the model specifications and parameter settings in Section 3.1, the Schelling model is simulated for μ ranging from 0.1 to 0.8, in increments of 0.1 with 10 replications for each parameter value. Figure 1 shows the impact of changing the tolerance parameter on the resulting level of segregation as measured by the average fraction of similar neighbours. For low intolerance values ($\mu = 0.1$), everyone is satisfied with their initial placement, resulting in very low segregation. However, even a slight increase to $\mu = 0.3$ leads to high levels of segregation of more than 0.7. Furthermore, as μ approaches 0.6, the level of segregation increases significantly, resulting in a fully segregated system. For satisfaction thresholds of 0.7 and higher, a huge decrease in segregation can be seen. This is because agents have difficulty finding satisfactory locations. This results in agents continually moving to new locations. At a certain point, there is no solution that can satisfy all (or a sufficient number of) agents. It is difficult to achieve a convergence with such a high threshold, and only a few of the conditions may satisfy all agents. This means that the resulting level of segregation is close to 0.5 as they stay in a different, but close to random configuration every time step (non-equilibrium dynamics). For 0.1-0.7 the model eventually reaches a stable state and for 0.7-0.76 there is a steep decrease with high variance.

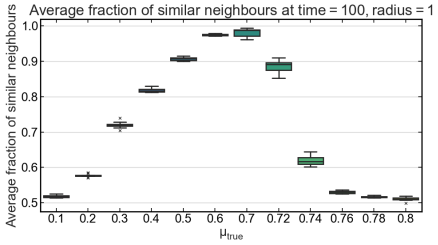

Fig. 1. Average fraction of similar neighbours at $t = 100$ as a function of the tolerance parameter (μ_{true}) and 10 replications for each μ_{true} value. Note that between 0.7-0.8, more values are added to include the high-variance region.

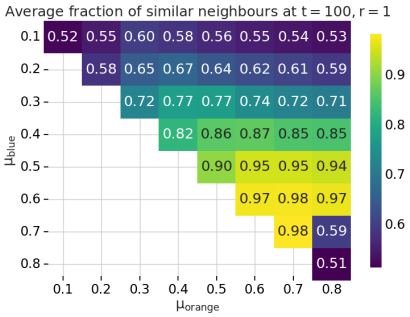

Fig. 2. Average fraction of similar neighbours at the end of a Schelling model run as a function of the satisfaction thresholds (μ_{blue} and μ_{orange}). The numbers are based on the average of 10 model runs for each combination.

4.1 Empirical calibration with spatio-temporal summary statistics

Figure 3 displays the posterior samples obtained using the NPE and NRE methods for different summary statistics. Ideally, NPE and NRE draw posterior samples close to the true tolerance value μ_{true} denoted by the dashed line. Although

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

all posteriors assign at least some probability mass around the true value, numerous posteriors are bimodal. This is most evident when μ_{true} is 0.1 or 0.74 and higher. In these cases, some posterior mass is centred around the true value, but most have two modes, one close to 0-0.1 and the other close to 0.8. This seems especially to be the case for the scalar summary statistics. This makes sense given its limited information, since the average fraction of similar neighbours is the same for a satisfaction threshold of 0.1 and 0.8 (Figure 1). Adding spatial or temporal information to the summary statistics makes this problem less severe, but not necessarily in all cases. For $\mu_{true} \in \{0.1, 0.8\}$, adding spatial information seems to lead to a more bimodal posterior than for the scalar summary statistic, for example. However, when adding spatial and temporal information simultaneously, the bi-modality disappears, and all posterior mass is centred around the true values.

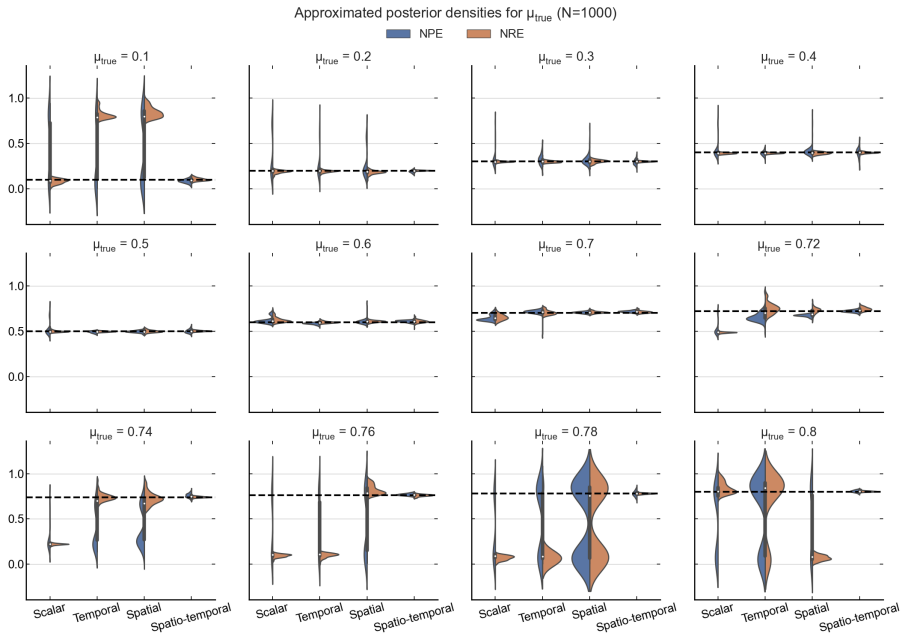

Fig. 3. Approximated posterior distributions for different summary statistics. Plots are based on 100 samples, dashed lines are the μ_{true} values.

To provide a more quantitative assessment of the methods, the different summary statistics and how they compare to ABC, the RMSE is reported in Figure 4. Moreover, because the Schelling model and the training of the neural networks contain stochastic elements, the calibration procedure is performed 10 times to

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

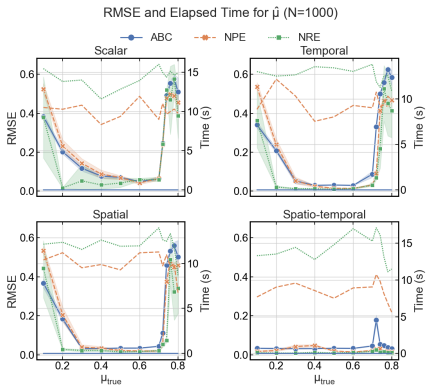

Fig. 4. RMSE for ABC, NPE, NRE and different summary statistics. For every μ value, the methods are trained 10 times, with a different training set consisting of 1000 samples. Hence, the reported RMSE is an average of 10 RMSEs, each based on 100 samples from the approximated posterior. Average elapsed time (lines without markers) is on the axis on the right.

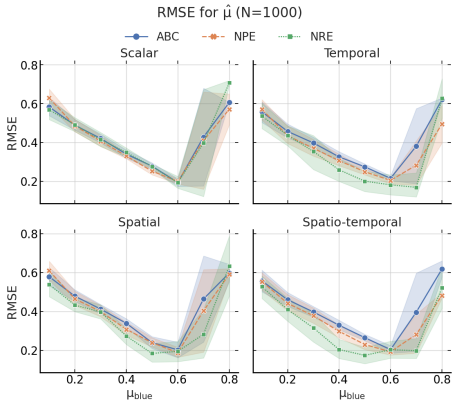

Fig. 5. RMSE when μ_{blue} is varied along the x-axis and averaged over the different values for μ_{orange}. The numbers are based on the average of 10 RMSEs for each combination.

calculate an average RMSE. Calculations are based on sampling 10% of the training budget, i.e., $M = 100$ samples for $N = 1000$.

In Figure 4 it can be seen that in the region $\mu_{true} \in [0.3, 0.7]$, the methods are able to estimate the value of μ_{true} quite accurately regardless of the summary statistic used, with most RMSEs below 0.10. The scalar summary statistic which contains only the average fraction of similar neighbours, seems to perform marginally worse than the other summary statistics in the easy-to-infer region (0.3-0.7). Hence, spatial and temporal trajectories provide important information on the true value of μ. Moreover, in most cases, NRE performs better than NPE and ABC. Using spatio-temporal information increases performance substantially, and NRE does better than NPE, while both outperform ABC. However, for NPE and $\mu_{true} \in \{0.3, 0.4, 0.5\}$ there is an increase in RMSE. In the analysis of Figure 3, the posteriors are wider than for NRE, but from only this figure it is not clear why this performance decrease is observed for NPE.

Moving towards the more difficult to infer regions of $\mu_{true} \leq 0.2$ and $\mu_{true} > 0.7$, one can see that the methods experience a significant increase in RMSE (Figure 4) and thus a decrease in performance. Using Figure 3, the larger RMSEs can be explained because the methods have difficulty distinguishing the regions around 0.1 and 0.8. For the scalar summary statistic, this makes sense as they lead to similar values and adding spatial or temporal information does not seem to change this. Interestingly, the addition of both leads to a substantial improvement, resulting in very low RMSEs. When $\mu_{true} = 0.2$, NRE performs

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

significantly better than NPE and ABC for all summary statistics. As mentioned when describing the dynamics of the Schelling model, both regions are rather extreme cases. For low values, everyone is satisfied immediately, and for high intolerance, agents keep moving around randomly but are unsatisfied. Interestingly, only adding spatio-temporal information simultaneously seems to be able to grasp this correctly.

Between 0.7 and 0.8, a region with a high variance in the average fraction of similar neighbours can be observed (Figure 1). Here, all methods have the greatest difficulty in terms of RMSE. Not much can be said on the differences in performance of the methods for scalar, spatial, or temporal summary statistics. However, adding spatio-temporal information leads to a significant decrease in RMSE, and this extra information benefits NRE and NPE even more so than ABC. Here, the usefulness of the neural-inspired methods might stand out, as they are better able to learn the (more) complex relationship between the input values and summary statistics in this region of the parameter space. However, this comes at a computational cost. NPE needs roughly 10 seconds to train and sample (with no clear difference between summary statistics, see Figure 4), while ABC runs in an instant. Moreover, NRE needs between 10 and 15 seconds, which is likely due to the extra MCMC step.

As these results can depend on the number of training samples used, the same experiment is repeated for different training budgets ($N \in \{500, 2000\}$). Figures A1-A2 show results that are very similar to those using $N = 1000$, but there are some noticeable differences. Firstly, in terms of computational cost, NPE and NRE take a couple of seconds less for $N = 500$, but significantly more time for $N = 2000$ (between 200 and 600 seconds). Moreover, using spatial summary statistics instead of scalar is not always beneficial. In the case of 2000 training samples, NRE performs even better for the scalar summary statistic than for temporal and spatial statistics separately in the case of $\mu_{true} \leq 0.2$. Hence, information redundancy can also hurt performance. Lastly, it seems that increasing or decreasing the training budget causes a slight general drop ($N = 2000$) or an increase ($N = 500$) in RMSE.

4.2 Multi-parameter problems: two groups with different thresholds

To test the methods on potentially more challenging and realistic problems, the Schelling model is extended with additional parameters. In this extension, groups have independent and possibly different satisfaction thresholds $(\mu_{blue}, \mu_{orange})$, where one group can be tolerant while the other is not, for example. In this case, calibration methods need to approximate the joint posterior distribution and a potentially more complex interaction pattern.

Figure 2 shows the average fraction of similar neighbours for different combinations of the two parameters. Almost complete segregation is the result if both groups have a threshold value between 0.4 and 0.7. If one or both groups have a low threshold value, segregation is low. In the case of both having high threshold, the agents are never satisfied, and hence, segregation is also low, as in

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

the homogeneous case (agents keep on moving). To approximate the posteriors, the methods are again given $N = 1000$ samples for training, and 100 values are sampled from the posteriors to calculate the RMSE with respect to the true values. This is repeated 10 times to arrive at an average RMSE per method and summary statistic, for each combination of the parameters.

For scalar summary statistics, the methods show similar performance (Figure 5). However, adding extra information does benefit NPE and NRE more than ABC. These results are in line with the one-parameter case. Note that in the figure, μ_{blue} is varied along the x-axis and that these numbers are averaged over the various μ_{orange} values. Furthermore, NPE and NRE sometimes perform more than 50%-80% better than ABC (plots are available upon request), which appears to be mainly in the moderate-high segregation region and when both μ values are large (> 0.7). Although ABC sometimes performs better than both NPE and NRE as well. In general, one could say the latter two are obtaining lower RMSEs in most of the parameter space, especially when increasing the amount of information provided by the summary statistics.

Furthermore, when the standard deviations of the tolerance thresholds are allowed to vary ($\sigma_{blue}, \sigma_{orange}$), the number of parameters goes from two to four. The standard deviations allow for control of the degree of heterogeneity within the group. Figure 6) shows three calibrations for different values of the four parameters. As four parameters must now be estimated, the sample size is set to 10,000. Although all methods perform better than the RMSE calculated on the prior values (except for two scalar summary statistic cases), NPE and NRE outperform ABC in almost all cases. Again, this may hint at the fact that these neural network methods are better able to learn in higher-dimensional spaces and/or more complex relationships. In addition, information redundancy sometimes degrades performance. In the last case for NRE, the scalar summary statistic has a lower RMSE than the others and in the first case, both temporal and spatial information separately have lower RMSEs for NPE and NRE.

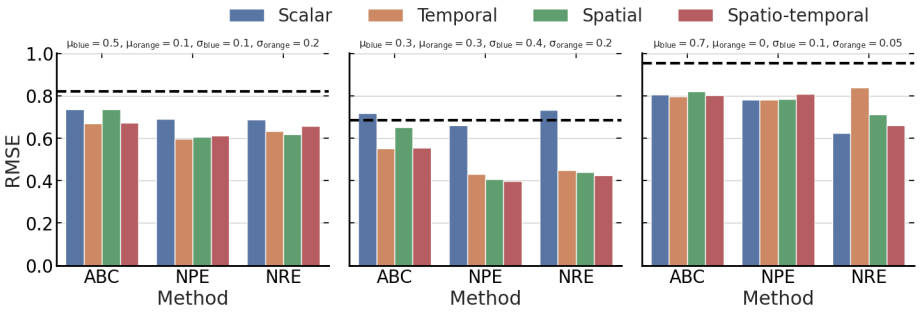

Fig. 6. RMSE of ABC, NPE and NRE when trained on a four parameter problem with 10,000 samples. The dashed line is the RMSE when using the prior values (i.e., training data) as estimates. Contrary to the other experiments, the RMSEs are for one training dataset only.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

5 Conclusion

In most ABMs, it is often hard to perform traditional statistical inference, as the likelihood is analytically intractable or computationally expensive to sample from. Fortunately, methods from the field of SBI, do not rely on the likelihood function to conduct parameter inference. This study assessed the performance of two recently developed SBI techniques using neural networks, NPE [6] and NRE [11], and compared them to a more commonly used one: ABC. In addition, we proposed a general approach, suitable for many ABMs, to provide spatio-temporal information in the summary statistics, which is a necessary and crucial ingredient for calibration. Compared with previous studies [5,4], the methods were evaluated for a large part of the parameter space of the satisfaction threshold (i.e., tolerance parameter) in Schelling's model of residential segregation [13], instead of only one particular set of parameter values with relatively straightforward dynamics. This includes regions of high variance in output and non-equilibrium, as well as equilibrium dynamics. This could be (more) challenging dynamics to perform empirical calibration on. In addition, changes in the amount of spatial and/or temporal information in the summary statistics, altering training budgets, and increasing the number of parameters to be estimated were tested.

In general, NRE showed better performance in terms of RMSE than ABC and NPE in most regions of the parameter space, regardless of which summary statistics were used. However, especially when adding the most elaborate spatio-temporal information, a clear performance increase could be seen between NRE and NPE versus ABC, but even more so for NRE than NPE. These conclusions remain unchanged when the training budget was decreased or increased. Having to calibrate two or four parameters instead of one also led to NRE outperforming NPE, and both surpassed ABC, in most cases. This improved performance may be due to the difference in the principal ideas behind the methods. Firstly, calibration methods using neural networks might be better able to learn (complex) relationships in the data than the ABC method and being more efficient. However, neural networks take considerably longer to train and sample, which can be of importance when selecting a method. Additionally, NRE transforms the approximation of the posterior into a classification problem. This may be an easier (supervised) learning task compared to the (unsupervised) learning task of the direct posterior approximation used by NPE, which may explain the improved performance of NRE over NPE. Notable is the increase in RMSE ($\mu \in \{0.3, 0.4, 0.5\}$) when spatio-temporal information is used with NPE. In the same region, NRE obtains lower RMSEs for using only spatial or temporal information. Moreover, this difference does not disappear when the training budget is increased, suggesting that the problem is not due to an increase in dimensionality. The persistence of this result over multiple training budgets hints at a possible different, as yet unidentified, cause.

Most ABMs have more parameters to calibrate and slower runtimes than the Schelling model. In such cases, the performance of the calibration methods becomes even more crucial. Changing the hyper-parameters of the neural net-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

works and using different architectures can improve performance compared to the default architectures used here. Assessing the performance when lowering the training budgets even more, as some models cannot be evaluated that many times, and increasing the number of parameters and/or summary statistics used (i.e., making the estimation problem more difficult) are interesting directions for future studies. One would expect a more rapid degradation in the performance of ABC, compared to NPE and NRE, because it throws away (many) samples. Additionally, in this specific setup, adding more information in the summary statistics is often beneficial, but for other problems, this might be different or not feasible to test various different summary statistics. Finding optimal high-level descriptions to capture the fundamental emergent patterns in the model through summary statistics and ensure the parameters are structurally identifiable could prove crucial in cases where systems are governed by more complex behaviour [17]. Moreover, all three methods have several extensions that can improve their performance. Sequential versions, which might improve efficiency but lose amortisation [5], could affect performance in different ways (i.e., some can be more efficient than others). In addition, learning summary statistics rather than hand-crafting them can improve performance [4]. However, a general problem with neural networks is that it is not clear how these methods learn, learning summary statistics would only worsen this problem. A summary statistic that could have potentially improved performance here, discriminating between $\mu_{true} = 0.1$ and $\mu_{true} = 0.8$, is the total number of relocations. For the former, the number will be low as agents will be satisfied quickly, and for the latter, it will be high. Lastly, the actual true value is used for performance assessment (which is often unknown for empirical data), but there are also several other assessment metrics. One could compare with the true posterior distribution if it is available or use posterior predictive checks and simulation-based calibration [5].

Our general approach (suitable for many ABMs) of providing spatio-temporal information combined with methods from SBI, most notably NRE, makes it possible to accurately estimate the posterior distributions of ABMs. This is even true in difficult regimes, where the model output exhibits a high sensitivity to parameters and has multiple potential solutions. This approach can be applied to real-world empirical observations that contain longitudinal information and/or data that can be aggregated at different spatial scales, paving the way for more realistic and empirically calibrated ABMs.

Acknowledgments. This study is part of the Computational Modelling of Primary School Segregation (COMPASS) project which is funded by the Dutch Inspectorate of Education and the City of Amsterdam.

Disclosure of Interests. The authors declare that they have no competing interests.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

References

1. Bonabeau, E.: Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the national academy of sciences **99**(suppl 3), 7280–7287 (2002)
2. Cranmer, K., Brehmer, J., Louppe, G.: The frontier of simulation-based inference. Proceedings of the National Academy of Sciences **117**(48), 30055–30062 (2020)
3. De Marchi, S., Page, S.E.: Agent-based models. Annual Review of political science **17**, 1–20 (2014)
4. Dyer, J., Cannon, P., Farmer, J.D., Schmon, S.M.: Calibrating agent-based models to microdata with graph neural networks. arXiv preprint arXiv:2206.07570 (2022)
5. Dyer, J., Cannon, P., Farmer, J.D., Schmon, S.M.: Black-box bayesian inference for agent-based models. Journal of Economic Dynamics and Control **161**, 104827 (2024)
6. Greenberg, D., Nonnenmacher, M., Macke, J.: Automatic posterior transformation for likelihood-free inference. In: International Conference on Machine Learning. pp. 2404–2414. PMLR (2019)
7. Hermans, J., Begy, V., Louppe, G.: Likelihood-free mcmc with amortized approximate ratio estimators. In: International conference on machine learning. pp. 4239–4248. PMLR (2020)
8. Lux, T.: Approximate bayesian inference for agent-based models in economics: a case study. Studies in Nonlinear Dynamics & Econometrics (0) (2022)
9. Marin, J.M., Pudlo, P., Robert, C.P., Ryder, R.J.: Approximate bayesian computational methods. Statistics and computing **22**(6), 1167–1180 (2012)
10. Massey, D.S., Denton, N.A.: The dimensions of residential segregation. Social forces **67**(2), 281–315 (1988)
11. Miller, B.K., Weniger, C., Forré, P.: Contrastive neural ratio estimation. Advances in Neural Information Processing Systems **35**, 3262–3278 (2022)
12. Papamakarios, G., Murray, I.: Fast ε-free inference of simulation models with bayesian conditional density estimation. Advances in neural information processing systems **29** (2016)
13. Schelling, T.C.: Dynamic models of segregation. Journal of mathematical sociology **1**(2), 143–186 (1971)
14. Sottoriva, A., Tavaré, S.: Integrating approximate bayesian computation with complex agent-based models for cancer research. In: Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers. pp. 57–66. Springer (2010)
15. Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.M., Durkan, C., Gonçalves, P.J., Greenberg, D.S., Macke, J.H.: Sbi–a toolkit for simulation-based inference. arXiv preprint arXiv:2007.09114 (2020)
16. Turner, B.M., Van Zandt, T.: A tutorial on approximate bayesian computation. Journal of Mathematical Psychology **56**(2), 69–85 (2012)
17. Wolpert, D.H., Grochow, J.A., Libby, E., DeDeo, S.: Optimal high-level descriptions of dynamical systems. arXiv preprint arXiv:1409.7403 (2014)
18. Zbair, M., Qaffou, A., Hilal, K.: Approximate bayesian estimation of parameters of an agent-based model in epidemiology. In: International Conference on Partial Differential Equations and Applications, Modeling and Simulation. pp. 302–314. Springer (2021)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18

A Appendix

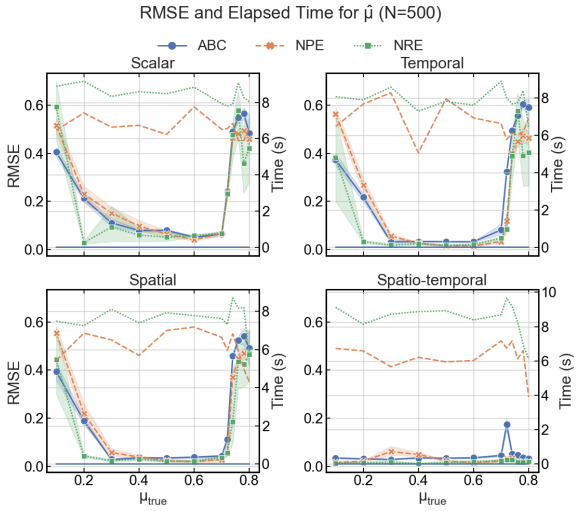

Fig. A1. RMSE for ABC, NPE, NRE and different summary statistics. For every μ value the methods are trained 10 times, with a different training set of 500 samples.

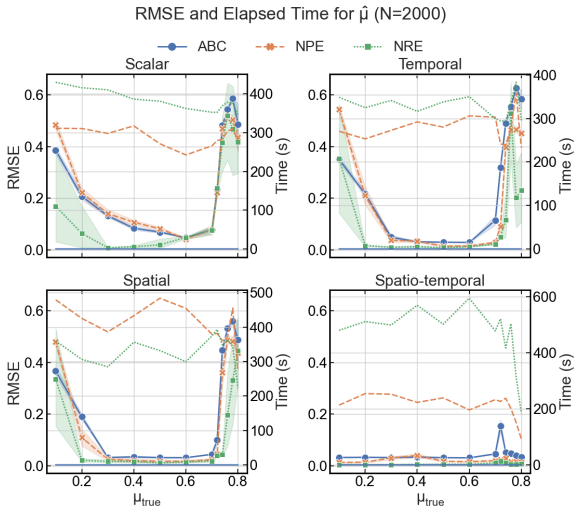

Fig. A2. RMSE for ABC, NPE, NRE and different summary statistics. For every μ value the methods are trained 10 times, with a different training set of 2000 samples.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_18