Chapter 14
Fixing Sample Biases in Experimental Data Using Agent-Based Modelling

Mike Farjam and Giangiacomo Bravo

Abstract We present how agent-based models can be used to correct for biases in a sample. The approach is generally useful for behavioural experiments where participants interact over time. The model we developed copied mechanics of a behavioural experiment conducted earlier, and agents in the model faced the same strategic choices as human participants did. We used the data from the experiment to calibrate agent behaviour such that agents reproduced patterns observed in the experiment. After this learning phase, we resampled agents such that their characteristics (political orientation) were similar to those found in the real world. We found that after the correction for the bias, agents produced patterns closer to those commonly found.

Keywords Agent-based modelling · Experiment · Bias · Methodology

14.1 Introduction

The goal of this work is to understand whether it is possible to use agent-based modelling to correct for biased samples in behavioural experiments. As interaction typically occurs over multiple rounds and participants receive feedback on and react to the behaviour of others, each individual does not represent an independent unit, which makes it impossible to correct for the bias by taking a stratified resample of the data originally collected. Different group compositions may indeed lead to very different dynamics, which cannot be captured by the resampling. We propose here to correct sample biases using an ABM where the agents' parameters are based on the data from the experiment, with the resampling that occurs at the agent level while the group dynamics is simulated by the model. By using data from a collective-risk

M. Farjam (✉) · G. Bravo
Centre for Data Intensive Sciences and Applications, Linnaeus University, Växjö, Sweden
e-mail: Mike.Farjam@lnu.se

H. Verhagen et al. (eds.), *Advances in Social Simulation*, Springer Proceedings in Complexity, https://doi.org/10.1007/978-3-030-34127-5_14

social dilemma (CRSD) experiment [1, 2], we show that a resampled ABM using more representative group compositions produces results that are closer to the ones from established research in the field than the outcome of the biased experiment.

14.2 The CRSD Experiment

Milinski et al.'s [3] CRSD has become the leading protocol to experimentally study climate change mitigation behaviour. Played in groups of six participants over ten rounds, in each round, participants decide whether to keep (part of) their endowments or to contribute to a "climate protection" account. At the end of the round, they receive feedback about the other group member's contributions. Money on the climate protection account is used to pay for pro-climate actions in the real world. If the total contribution of the group is above a fixed threshold, the money remaining in the private account is paid to each participant. If the total contribution is lower, all the private account money is lost with a certain probability.

More precisely, each participant is first endowed with 40 experimental currency units (ECU) and can contribute {0, 2, 4} ECU to the climate account in each round. The probability of losing the private account money if a threshold of 120 ECU in the climate protection account is not reached is referred to as p and is $\{0.1, 0.5, 0.9\}$. Milinski et al. showed that, using a sample of German students in a university lab, all groups in the $p = 0.1$ condition, 90% of the groups in the $p = 0.5$ one and 50% of the groups in the $p = 0.9$ one *failed* to successfully solve the dilemma. Similar findings were produced by subsequent lab experiments using similar participants' populations [2, 4].

Being interested in checking whether these results were robust to a change of the participants' population, we recently replicated Milinski's experiment using a sample of adult US citizens recruited through mTurk, a common recruitment tool for behavioural experiments [5]. To our surprise, we recorded much higher contribution levels, leading to only 19%, 10% and 7% of the groups failing to reach the threshold in the $p = 0.1$, $p = 0.5$ and $p = 0.9$ conditions, respectively. Deeper analyses showed that, while the age and gender distributions of the participants were more or less comparable to the US ones, our sample over-represented left-wing, environmental-concerned people. Both variables were measured on a 0–10 point scale in post-experimental questionnaires and showed medians of 4 (left wing 0–right wing 10) and 8 (low concern 0–high concern 10), respectively. The median values for similar questions in the US sample of the World Value Survey (WVS), wave 2010–2014, were 5.5 for both political orientation and environmental concern[1] (Fig. 14.1).

The importance of these variables is confirmed by the fact that our measure of environmental concern and, to a lower extend, of left-wing orientation significantly correlated with contributions ($r = 0.44$ and $r = 0.17$, respectively). In addition,

[1] To help the comparison, the original values were rescaled into our 11-point scale.

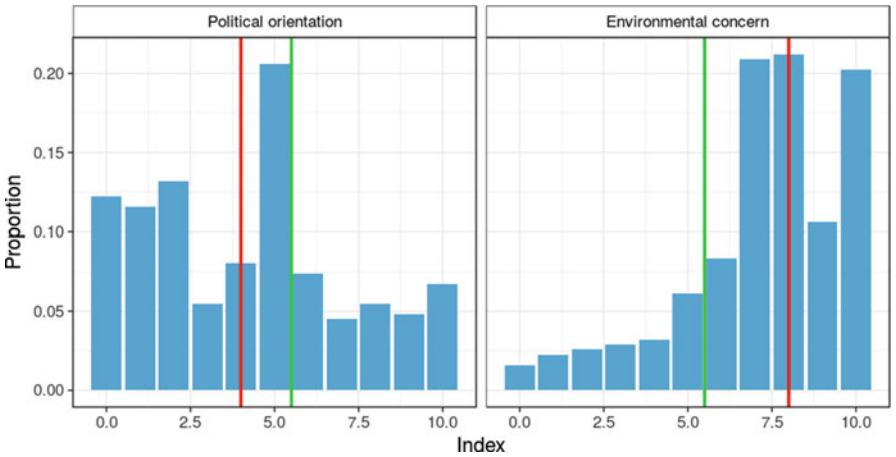

Fig. 14.1 Distribution of the political orientation and environmental concern scales among participants in the mTurk experiment. Median values in the experiment are indicated as red lines, actual USA values from the WVS as red lines

knowing that the contributed money was used to buy carbon compensation credits, some participants may have taken this opportunity to produce positive real-world externalities.[2] Also taking into account the strong political polarization of the topic in the USA, this probably led to a much higher willingness to contribute than in standard laboratory experiments.

14.3 Methods

We developed an ABM in NetLogo 6.0 [6] reproducing the CRSD game dynamics. Agents' data mapping the experiment participants are first uploaded in the model, and then the game goes on as in the original experiments. In each period, agents choose a level of contribution following the procedure described below. Then contributions are aggregated for each group, and agents are "informed" of the total group contribution and the remaining distance from the threshold. After 10 periods, corresponding to the 10 experimental rounds, the proportion of groups that reached the threshold is measured.

To determine the behaviour of agents starting from the empirical data, we estimated a linear mixed effects model. The political orientation and environmental concern scales were used to predict the contribution in the experiment in each period, along with the value of p used for that specific group, a measure of inequality in contributions (the ratio between the subject contribution and the mean

[2]Following the protocol, 800$ were used to offset the emission of 2050 t CO_2.

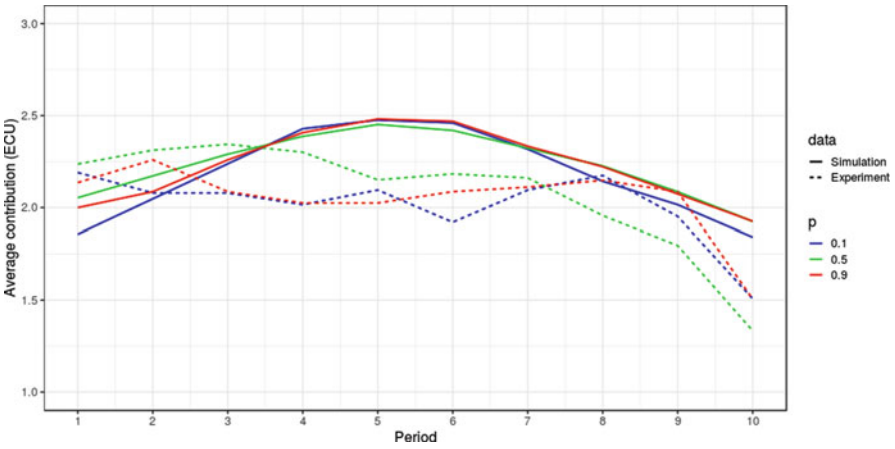

Fig. 14.2 Contribution dynamics in the empirical and simulated data

contribution of other group members in the previous period) and the distance from the threshold weighted by the period number. The resulting coefficients, along with the subject characteristics, such as political orientation and environmental concern, were then used to predict individual contributions. To transform the continuous output of the linear model into the discrete set of possible contributions, two cutting points were estimated by calibrating the model in order to maximize its capacity to predict the success of each group.

The resulting model was able to reproduce well the outcome of the experiment: predictions were correct for 70/70 groups that reached the threshold and 8/9 groups that did not. In addition, the model was able to capture the declining dynamics of contributions in the final part of the experiment, once most group passed the threshold, although it slightly over-estimated contributions in the central part of the game (Fig. 14.2).

To resample the agent population, we implemented a simple NetLogo procedure where one agent with a value of political orientation below the actual US median was randomly deleted and replaced by a copy of another agent having political orientation above the median. The procedure was repeated until the median in the simulation reached the actual US level. The same occurred for the environmental concern variable, except that agents with high concern were deleted and replaced by ones with low concern. Once completed the resampling procedure, the CRDS game was played as above, except that all groups were reshuffled. One hundred repetitions of the model were run for each level of p and the corresponding proportion of groups reaching the threshold recorded. On average, 38.4% of the groups in $p = 0.1$ failed to reach the threshold, 35.3% in $p = 0.5$ and 31.7% in $p = 0.9$, with a clear decrease of the success cases in comparison with the mTurk experiment results.

14.4 Conclusions

The outcome of our resampling model were much closer to the ones in Milinski's experiment [3], at least in the $p = 0.9$ condition. However, the same did not fully occur in the other two conditions condition where, although the model led to a significant decrease in comparison with the mTurk experiment, it still predicted that around 60% of the groups would pass the threshold, while none did so in Milinski's $p = 0.1$ condition and only 10% in the $p = 0.5$ one [3].

One reason for this could be that our model is fully data-driven and, as a consequence, has limits in correctly predicting behaviour that is too far from the one observed in the original sample. From this point of view, a model based on sounder cognitive mechanisms could improve the result. On the other hand, students often exhibit more selfish behaviour in experiments than older participants [7], which means that our model outcome could fit the actual behaviour of the general US population better than Milinski's one based on a sample of German students. In addition, subsequent experiments that reproduced or slightly modified Mikinski's design recorded higher contributions than the original study [4, 8]. For instance, Tavoni et al. [8], who replicated Milinski's $p = 0.5$ treatment, had 50% of the groups failing to reach the threshold: a clear decrease in comparison with Milinski and a result much closer to our resampled-group outcome.

While it is difficult to solve this issue without running another experiment on an appropriate sample, the method we illustrated in this paper could be more carefully tested by splitting existing experimental dataset into different samples and trying to use one part to predict the other. Future work will focus on doing this.

References

1. Farjam, M., Nikolaychuk, O., Bravo, G.: Does risk communication really decrease cooperation in climate change mitigation? Clim. Chang. **149**(2), 147–158 (2018)
2. Farjam, M., Nikolaychuk, O., Bravo, G.: Experimental evidence of an environmental attitude-behavior gap in high-cost situations. Ecol. Econ. **166**, 106–434 (2019)
3. Milinski, M., Sommerfeld, R.D., Krambeck, H.J., Reed, F.A., Marotzke, J.: The collective-risk social dilemma and the prevention of simulated dangerous climate change. Proc. Natl. Acad. Sci. **105**(7), 2291–2294 (2008)
4. Dannenberg, A., Löschel, A., Paolacci, G., Reif, C., Tavoni, A.: On the provision of public goods with probabilistic and ambiguous thresholds. Environ. Resour. Econ. **61**(3), 365–383 (2015)
5. Paolacci, G., Chandler, J., Ipeirotis, P.G.: Running experiments on Amazon Mechanical Turk. Judgm. Decis. Mak. **5**(5), 411–419 (2010)
6. Wilensky, U.: NetLogo. Center for Connected Learning and Computer-Based Modeling. Northwestern University, Evanston (1999)
7. Belot, M., Duch, R., Miller, L.: A comprehensive comparison of students and non-students in classic experimental games. J. Econ. Behav. Organ. **113**, 26–33 (2015)
8. Tavoni, A., Dannenberg, A., Kallis, G., Löschel, A.: Inequality, communication, and the avoidance of disastrous climate change in a public goods game. Proc. Natl. Acad. Sci. **108**(29), 11825–11829 (2011)