UNIVERSITÀ POLITECNICA DELLE MARCHE

Centro Studi Sulle Politiche Economiche, Rurali e Ambientali

associazioneAlessandroBartola
studi e ricerche di economia e politica agraria

Paper prepared for the 122nd EAAE Seminar

"EVIDENCE-BASED AGRICULTURAL AND RURAL POLICY MAKING: METHODOLOGICAL AND EMPIRICAL CHALLENGES OF POLICY EVALUATION"

Ancona, February 17-18, 2011

The sampling bias in multi-agent simulation models

Buysse, J.[1], Frija A.[1], Van der Straeten B.[1], Nolte S.[1], Lauwers L.[1,2], Claeys D.[2] and Van Huylenbroeck G.[1]

1 Department of agricultural economics, Ghent University, Coupure Links 653, B-9000 Gent, Belgium
2 Institute for Agricultural and Fisheries Research, Merelbeke, Belgium

J.Buysse@ugent.be

The sampling bias in multi-agent simulation models

Buysse J., Frija A., Van der Straeten B., Nolte S., Lauwers L., Claeys D. and Van Huylenbroeck G.

Abstract

For practical considerations, it is in some case impossible to simulate MAS models at population level. The current paper shows that MAS models applied to samples with heterogeneous costs of interactions between agents have biased results. Heterogeneous costs of interactions in MAS models can come from the spatial dimension in MAS models or from fixed costs per interaction. The paper presents two correction procedures to remove the sampling bias and to increase the reliability of the outcome. The correction procedures can be very promising for future applications of MAS models because it becomes possible to deploy more complex models without bias on more detailed datasets that are only available at sample level, which will be the case for country- or EU-wide MAS applications.

Keywords: MAS, bias, correction, resampling

JEL classification: Q12, Q18, Q51, Q52.

1. INTRODUCTION

The importance of taking the heterogeneity of responses into account, when modelling farmers decisions and policy distributional effects has increasingly been recognized.

Agent-based models (ABM) or Multi-agent simulation models (MAS) are a developed in order to investigate this heterogeneity and the distributional effects. MAS recognise the importance of the interplay occurring at two different scales of a given system: the macro structure and the micro structure. Many systems are characterized by the fact that their aggregate properties cannot be deduced simply by looking at how each component behaves, the interaction structure itself playing a crucial role (Leombruni and Richiardi, 2005)

Therefore, "MAS models have a one-to-one representation of real-world farm households and computational agents, which eliminates the need to define a limited number of representative farm households and makes MAS highly suitable for representing heterogeneity in both socioeconomic and biophysical terms" (Schreinemachers and Berger, 2006). Moreover, the simulation of interactions between farms enables the modeller to simulate in a realistic way spatial effects, transaction costs of exchange of production factors, propensity of innovations,…

Because of their conceptual framework, MAS models are used to tailor mathematical models to the real world situation and to capture the heterogeneity of opportunities and constraints at the individual level, their adaptive capacities, as well as to quantify the distributional effect change (Schreinemachers and Berger, 2006).

However, like any other approach, MAS has also certain drawbacks. An important pitfall of the one-to-one agent representation in an empirical MAS model is that the one-to-one agent interaction can only be perfectly represented in a full population model. In the case of MAS models applied to a sample of farms, the farms in the sample can not interact with their real-world interacting farms because they are not all represented in the sample. As a result, the one-to-one agent interaction simulated in the model can not completely reflect the real world interaction. The current paper illustrates that there is a systematic bias in the results in absence of a full population of agents if heterogenous costs of interactions between agents are simulated in the model.

The relevance of this problem is very high because i) most empirical models have to rely data based on a sample of farms and not the full population ii) the clear objective of MAS is to simulate interactions between agents, which are rarely homogenous among the population. Despite the relevance of the problem, there is, to our knowledge, no paper that quantifies the impact of the sampling bias and describes mechanism to deal with it. Yet, some of the MAS models address the issue of the sampling bias by artificially generating a full population dataset (Schreinemachers and Berger, 2006; Happe and Balmann, 2003). Other papers relax the model specification and simulate homogenous costs of interactions (Buysse et al., 2007). Finally, there are papers that do not specifically explain if and how the sampling bias is tackled (Möhring et al., 2010).

The first objective of the current paper is to test and to illustrate the impact of sampling and the sampling biases which results if there are heterogeneous transactions costs between agents. The second objective is to discuss and to develop mechanisms that can remove this sampling bias.

Therefore, the remaining of the paper is as follows. The next section start with a literature review of MAS models in agriculture and why and which type of data is used. The third section analyses the sampling bias more in detail using a pulished empirical model that until now runs on a full population and a small illustrate model applied on an artificial population of 500 farms. The model is used to quantify the sampling bias and to illustrate corrections that can deal with the sampling bias. The paper concludes with a discussion and conclusion.

2. LITERATURE OVERVIEW OF TYPES OF DATA IN MAS MODELS

Möhring et al. (2010) have made already an overview of various methods for defining agents and generating the agent population in MAS models.

One type of approach that was not mentioned in Möhring et al. (2010) is the specification of MAS models at population level and using on administrative datasets or census data such as Van der Straeten et al. (2010). The main advantages of the approach is that the full heterogeneity in the data is represented and there is no sampling bias. With increasing access to High Performance Computation (HPC) systems and with a model of limited computational complexity, Van der Straeten et al. (2010) has show to be able to run simulations for a population of 30000 agents. Censuses do capture all farm households in the study area but, for financial and time constraints, cannot provide in-depth data of high quality (Berger and Schreinemachers (2006). Therefore, other empirical applications of MAS use a smaller regions where detailed information on a limited sample can be gathered. Two of these application are cited in Möhring et al. (2010): “Lauber (2006) and Albisser (2008), for example, have described

Swiss communities with 72 and 30 existing farms, respectively; but the results of such case studies can only be generalised to a limited extent."

The lack of high in-depth data of administrative datasets and the lack of representativity of case studies have motivated researchers to search for approaches to use samples for data collection for MAS models. Two approaches have been described to generate a full population based on sampled information.

Happe and Balmann (2003) use a sample of 12 farms that are defined as typical regional farms. The full population is generated by assigning a frequency to each farm type which is the number of times this particular farming system is represented in the region. Through identical multiplication ('cloning') of the farms – as a function of their occurrence in the population – an agent population is generated which corresponds to the actual size of the region.

This approach does not represent the actual heterogeneity of the individual farms. Therefore, more sophisticated methods for defining agents and generating the agent population have been developed (Möhring et al., 2010).

Berger and Schreinemachers (2006) use a Monte Carlo procedure to produce a full population that also represents the full heterogeneity. The advantage of the approach is that different data sources can be used that are not all available at population level. This is a very good approach for correctly simulating at population level if not all data is available for the complete population. The disadvantage of the approach is that the dataset used in the model becomes very big if a large region is to be modelled. For smaller countries such a Switzerland (The Swissland model of Möhring) or regions such a Flanders (the model of manure allocation of the authors of this article) the amount of agents in the model would be more than 30000. For complex models with non-linear behaviour and/or mixed integers models, it is currently not possible yet to build an operational model of this size.

Therefore, teams working on the Swissland model (described in Möhring et al., 2010) and also the authors of this article are exploring to run MAS models that are representative for a bigger region but only simulate the sample of farms instead of the complete population.

As long as there are no heterogeneous costs of interactions between agents, the sampling would not be problematic. Unfortunately, the following sections of the paper prove that sample-based MAS models with spatial interactions between agents or with interactions with fixed costs have biased outcomes.

The final section of the paper will present methods to correct for the sampling bias. The presented approach offers a practical solution for MAS models that are impossible to apply at a full population. In addition, the analysis of the bias gives also insights of the benefits of corrective procedures to improve the representativeness of the models.

3. ANALYSIS OF THE SAMPLING BIAS IN MAS

3.1. The origin of the sampling bias

The sampling bias in MAS models is caused by the fact that interactions between agents change if agents are removed from the model, which is essentially what sampling does. The model of Van der Straeten et al. (2010) can illustrate very clearly how this happens. The model of Van der Straeten et al. (2010) deals with manure disposal and optimises for each agent its manure disposal costs by calculating the shortest distance to transport manure to other farms. The reduction of the sample size on which the model is applied will increase the average distance of the closest neighbouring farm where excess manure can be transported to. The result is that one can expect increasing simulated costs with decreasing sample sizes due to the increased distance. The quantification of this effect is illustrated in next subsection.

Another source of the sampling bias is the presence of heterogeneous supply and a fixed cost of interactions between agents as illustrated in Buysse et al. (2010). Buysse et al. (2010) simulate bilateral trade of production rights between farms with a fixed cost per transaction. The probability that an ideal amount of rights is available at another farm will decrease with a decreasing sample size. In other words, the sample size again determines the cost of interactions between farms. The application of the model in Buysse et al. (2010) was intentionally applied to a smaller sample to simulate the imperfect information in the exchange of rights. However, one should be careful about implementing fixed costs on interactions to prevent that sampling has an unintentional bias on the model results.

3.2. Illustration of the sampling bias with an operational model

The illustration with an operational model is based on the case of Van der Straeten et al. (2010) where the full population is known. The same model has been applied on smaller samples that are 100 times randomly resampled. The sampling bias can be derived from comparing the average cost and its standard deviation of the resampling procedure for different sample sizes (100; 200; 500 and 750 farmers).

The results simply confirm that a bias exists and, in line with theoretical expectations, that it decreases when the sample size increases. For a subsample of 100 farmers the average cost is 52% over the average cost calculated using the full population data while for a subsample of 750 farmers, the average cost per farmer is only 27% of the full population average (Table 1).

Table 1: Mean and standard deviation of average cost estimates for different bootstrapping simulations

Bootstrap	Number of repetitions	Average cost simulated	SD	Average cost simulated/Average cost of population
S= 100 (0.26%)	100	3210.88	1200.77	+ 59 %
S= 200 (0.52%)	100	2698.58	666.24	+ 33 %
S= 500 (1.31%)	100	2581.09	755.91	+ 28 %
S= 750 (2%)	100	2571.13	536.30	+ 27 %
Full population (100%)	-	2016.49	-	-

S= Sample size (percentage of sample size compared to full population)

The next section of the paper explores methods for solving the problem by using a simplified model applied to a synthetic population of 500 farms.

3.3. Illustration of the sampling bias with a simplified model

The simplified model has similar features of the model of Van der Straeten et al. (2010). The model minimises the transport costs of emissions and the cost of emission abatement. The model specification in algebraic notation with variables in Greek letters and parameters in Latin letter is as follows:

Minimize $\Sigma_n (\Sigma_m c_{nm} \tau_{nm} + \omega_n p)$

s.t.

$e_n + \Sigma_m \tau_{mn} - \Sigma_m \tau_{nm} \leq r_n + \omega_n$

where

n and m are farm indices,

τ_{nm} is the amount of transported emission form n to m,

ω_n is the amount of emission abatement of agent n,

e_n is the amount of emission of farm n,

r_n is the amount of emission rights of farm n,

c_{nm} is the transport cost per transported emission from farm n to m ,

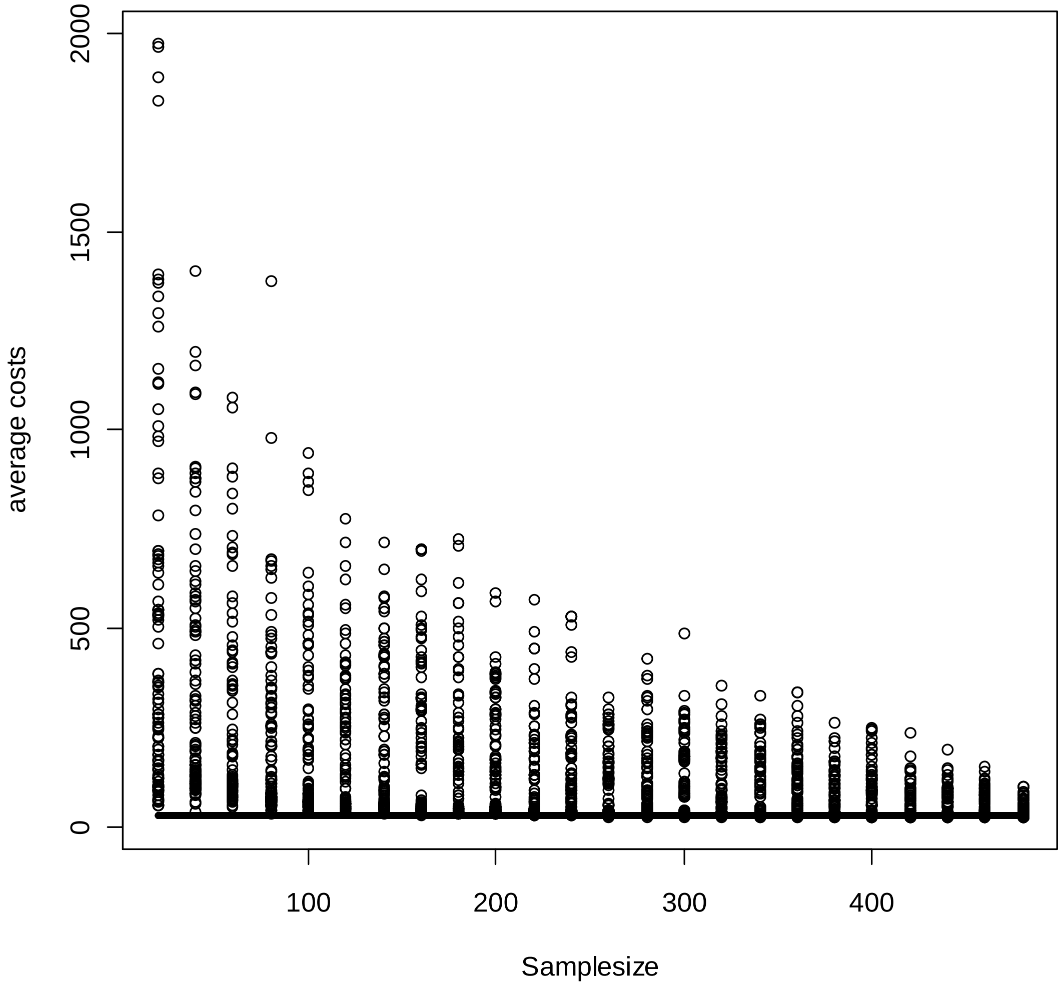
p is the penalty per overused emission right.

The synthetic dataset of 500 agent is generated by random assignment of the parameters e_n, r_n and c_{nm}. The values of e_n and r_n are selected from a random distribution with an average of 100 and standard deviation of 20. The values of c_{nm} are selected from a uniform distribution between 0 and 200. The emission abatement costs ‘p’ is 150.

The solution of the model to the complete sample results in a average cost per farm of 27.74. The sampling bias can again be illustrated using a resampling procedure as applied to the operational MAS model.

The results of this resampling procedure are illustrated in the 0 where each point represents the result of an optimisation of a resampled dataset at different sample sizes. The black straight line indicates the average costs obtained from running the model on the population.

Figure 1: The simulated average costs as a function of the selected sample size



0 shows some very interesting results. First, it is clear that the order of magnitude of the sampling bias can be very large. This is the case in the simplified model because only the heterogeneous transactions costs are simulated. Agent specific costs that are not affected by the sampling are not included in the model. Nevertheless, the simplified model illustrates the importance of not ignoring possible sampling biases in MAS models: the simulated costs can increase with a factor 10 for a sample size of 5% of the population compared with a simulation on the complete population.

Another remarkable result is the nonlinear effect of the sample size on the bias and the fact that the variation is very large. Even the 100 samples of a sample size of 480 have an average simulated cost of 39, which is 33% more than the population average. This very large difference can be explained by the fact that the subsamples do not always satisfy the population balance (the macrobalance). In the population, the amount of emission is smaller than the total amount of emission rights (Σ_n e_n < Σ_n r_n). This macrobalance is not for every subsample satisfied (Σ_n e_n > Σ_n r_n) resulting in a large upward bias of the costs because the abatement costs (ω_n p) are much larger than the transport costs of emissions (Σ_m c_{nm} τ_{nm}).

This observation has motivated our first type of correction for the sampling bias: a macrobalance correction which is discussed in the next subsection.

3.4. Illustration of a macrobalance correction on the sampling bias with a simplified model

The same model and resampling procedure are applied as illustrated before but for each subsample a correction factor is applied to e_n to make sure that the macrobalance of the population (Σ_n e_n / Σ_n r_n) also holds at the sample level. The results of this resampling procedure are illustrated in 0.

Figure 2: The simulated average costs with a macrobalance correction as a function of the selected sample size

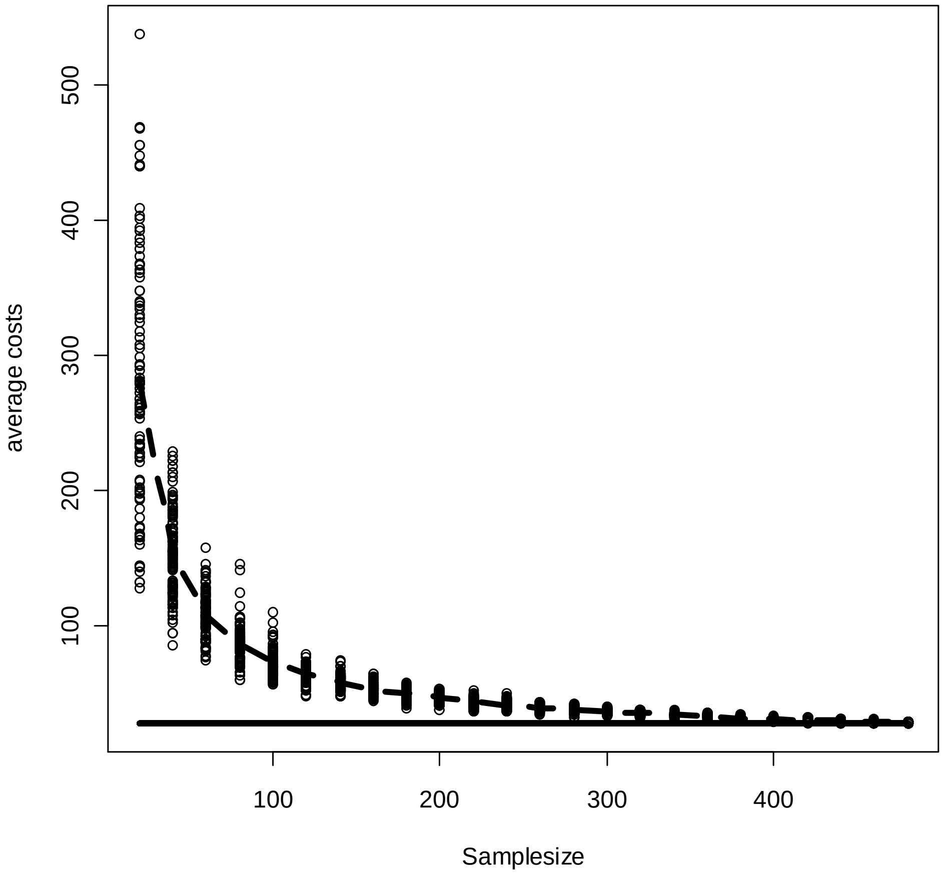


Figure 2 clearly illustrates the benefit of the macrobalance correction because both the bias and the variation are reduced significantly. Yet, it is clear that an important bias remains in the simulation. This bias is due to the reduced interaction opportunities in smaller samples, which is in our model reflected by the increased distance between agents. If one can quantify the bias, it is also possible to calculate a correction factor that can remove the bias caused by the increased distance. Therefore, we have estimated a function that explains the simulated costs as a function of the sample size. The dashed line in 0 is a polynomial of degree 10 where the average costs is fitted as a function of the sample size. The coefficients of this polynomial are shown in Table 30.

Table 3: The coefficients of the polynomial of the simulated average costs on the sample size

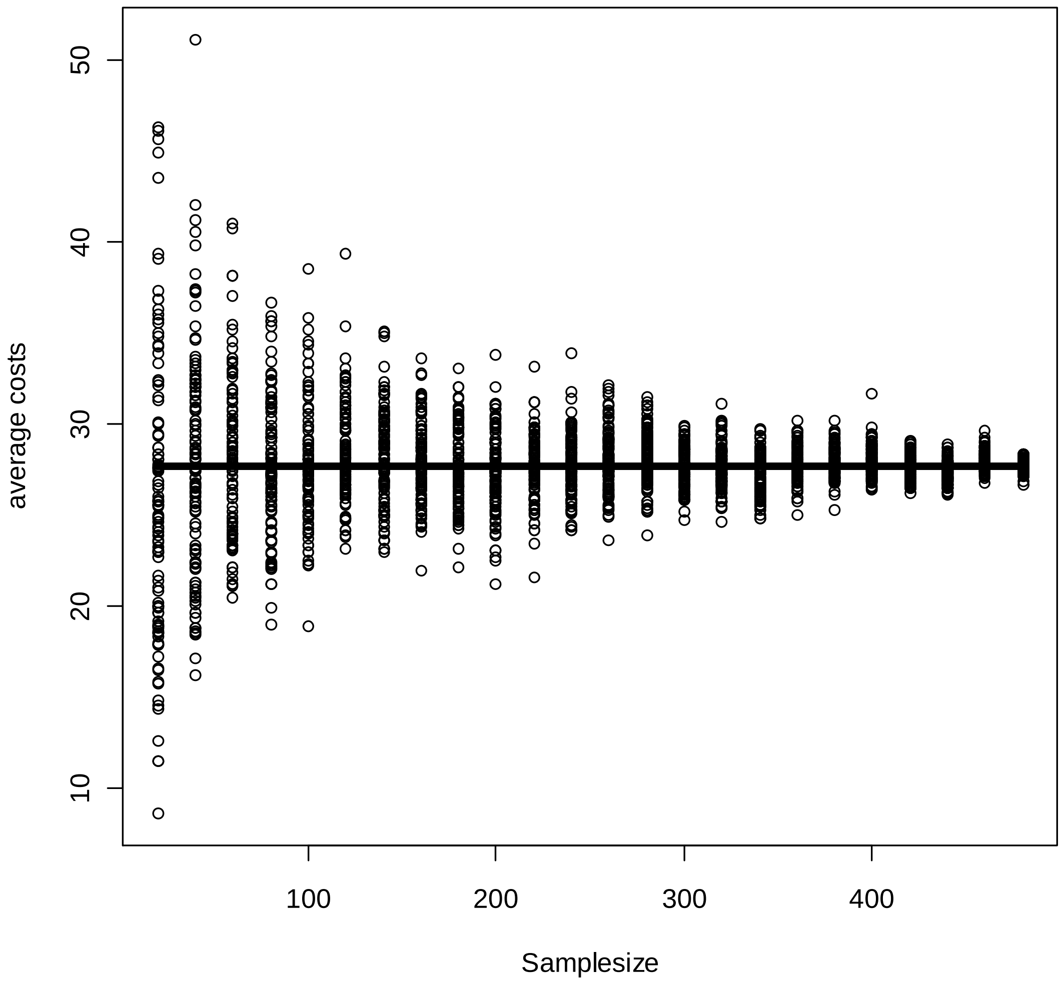
	Estimate	Std. Error	t value	Pr(>|t|)	
(sample size)^0	6,11E+05	2,68E+04	22.754	< 2e-16	***
(sample size)^1	-2,53E+04	2,62E+03	-9.657	< 2e-16	***
(sample size)^2	5,64E+02	9,35E+01	6.032	1.88e-09	***
(sample size)^3	-7,46E+00	1,69E+00	-4.403	1.11e-05	***
(sample size)^4	6,24E-02	1,78E-02	3.505	0.000464	***
(sample size)^5	-3,42E-04	1,16E-04	-2.944	0.003267	**
(sample size)^6	1,24E-06	4,84E-07	2.563	0.010450	*
(sample size)^7	-2,95E-09	1,29E-09	-2.286	0.022337	*
(sample size)^8	4,42E-12	2,13E-12	2.076	0.037970	*
(sample size)^9	-3,78E-15	1,98E-15	-1.911	0.056064	.
(sample size)^10	1,41E-18	7,91E-19	1.778	0.075497	.

These fitted values of the estimated polynomial are in the next section used to apply a correction factor on the distance.

3.5. Illustration of a distance correction on the sampling bias with a simplified model

The second correction for the sampling bias is based on the fitted values of the regression in 0. The coefficient allow to calculate a correction factor as a function of the sample size. The distance (c_{nm}) is in every simulation divided by this correction factor and multiplied by the population average costs. The same model and resampling procedure are applied as illustrated earlier. The results of this resampling procedure are illustrated in Figure 3.

Figure 3: The simulated average costs with a macrobalance and a distance correction as a function of the selected sample size



The results in 0 illustrate that the systematic bias is removed by the combination of the two corrections. It is made to the attention of the reader that only the procedure to calculate the correction factor can be generically applied on different models or model types. The correction factor itself is obviously model specific.

4. Discussion

The presented correction procedures are very promising for our own future applications of our MAS models. Currently, we have the model of Van der Straeten et al. (2010) applied on population data. However, we want to refine the model by using additional data available from the Farm Accountancy Data Network. In addition, the model of Van der Straeten et al. (2010)

does not simulate fixed transaction costs and nonlinear effects which would it make impossible otherwise to simulate the interactions between more than 30000 agents in MINLP (mixed integer non linear programming) models.

The possibility to calculate a macrobalance correction and a distance correction would allow us to deploy a more complex model on the FADN sample of 600 agents without the systematic sampling bias. We are convinced that this solution will also be useful for other MAS models because more detailed datasets become available and the complexity of the simulated agent decision increase.

However, in some cases it will be impossible to derive the correction factor because the population data are not available. Therefore, future research should try to assign correction factors based on information directly available in the sample. In the case of the presented simplified model, Table 4 compares the minimum interaction costs to the neighbouring farms (Σ_n (Min$_m$ c_{nm}) / Σ_n 1) that can be directly observed in the sample with the estimated average costs after running the resampling procedure. The comparison shows a strong correlation between the two and their related correction factors. Therefore, it would be possible to calculate the correction factor based on the observed minimum costs if it is impossible to run the resampling procedure.

Table 4: A comparison of the minimum interaction cost and the distance correction

Sample size	Σ_n (Min$_m$ c_{nm}) / Σ_n 1	Correction based on observed minimum cost	Estimated average cost	Correction based on estimated average cost
20	11.17	8.09	280.41	10.11
40	6.02	4.36	154.64	5.57
60	4.31	3.12	107.48	3.87
80	3.53	2.56	86.48	3.12
100	2.94	2.13	73.52	2.65
120	2.62	1.90	64.01	2.31
140	2.40	1.74	57.21	2.06
160	2.22	1.61	52.61	1.90
180	2.08	1.51	49.35	1.78
200	2.00	1.45	46.56	1.68
220	1.89	1.37	43.82	1.58
240	1.82	1.32	41.16	1.48
260	1.74	1.26	38.89	1.40
280	1.70	1.23	37.24	1.34
300	1.65	1.20	36.15	1.30
320	1.62	1.17	35.28	1.27
340	1.57	1.14	34.24	1.23
360	1.54	1.12	32.90	1.19
380	1.51	1.10	31.53	1.14
400	1.48	1.08	30.63	1.10
420	1.47	1.06	30.36	1.09
440	1.44	1.05	30.00	1.08
460	1.42	1.03	28.56	1.03
480	1.40	1.02	28.40	1.02
500	1.38	1.00	27.74	1.00

5. CONCLUSIONS

For practical considerations, it is in some case impossible to simulate MAS models at population level. The current paper has shown that MAS models applied to samples with heterogeneous costs of interactions between agents have biased results. Heterogeneous costs of interactions in MAS models can come from the spatial dimension in MAS models or from fixed costs per interactions.

The paper has presented two correction procedures to remove the sampling bias and to increase the reliability of the outcome. The correction procedures can be very promising for future applications of MAS models because it becomes possible to deploy more complex models on more detailed datasets that are only available at sample level without the sampling bias.

Berger and Schreinemacher (2006) solve the sampling bias by generating a population with a monte carlo method from data at sample level. This is a valid alternative but has the disadvantage of running the simulations on a very large dataset. For country- or EU-wide MAS application this is not always possible.

ACKNOWLEDGMENT

This research was funded by the institute for the promotion of innovation by science and technology in Flanders (IWT-Vlaanderen).

REFERENCES

Berger, T. (2001): Agent-based Spatial Models Applied to Agriculture: A Simulation Tool for Technology Diffusion, Resource Use Changes and Policy Analysis. Agricultural Economics, 25, 2, 1-16.

Berger, T. and P. Schreinemachers (2006): Creating agents and landscapes for multiagent systems from random samples. Ecology and Society, 11, 2, Art. 19.

Buysse, J., B. Fernagut, O. Harmignie, B. Henry de Frahan, L. Lauwers, P. Polomé, G. Van Huylenbroeck and J. Van Meensel (2007): Farm-based modelling of the EU sugar reform: impact on Belgian Sugar beet suppliers. *European Review of Agricultural Economics* 34 (1): 21-52.

Happe, K., and A., Balmann. (2003): Structural, Efficiency And Income Effects Of Direct Payments - An Agent-Based Analysis Of Alternative Payment Schemes For The German Region Of Hohenlohe. Meeting International Association of Agricultural Economists, August 16-22, 2003, Durban, South Africa.

Van der Straeten, B., J. Buysse, S. Nolte, L. Lauwers, D. Claeys and G. Van Huylenbroeck (2010): A multi-agent simulation model for spatial optimisation of manure allocation. *Journal of environmental planning and management* 53 (8): 1011-1030.

Van der Straeten, B., J. Buysse, S. Nolte, L. Lauwers, D. Claeys and G. Van Huylenbroeck (2011): Markets of concentration permits: the case of manure policy. *Journal of Environmental Economics and Management* (submitted).

Leombruni, R. and M. Richiardi. (2005): Why are economists sceptical about agent-based simulations? Physica A 355 (2005): 103–109.

Möhring, A., A., Zimmerman, G., Mack, S., MANN, A., Ferjani, M., Gennaio. (2010): Modellig structural change in the agricultural sector- an agent-based approach using FADN data from individual farms, paper presented at the 114th EAAE Seminar 'Structural Change in Agricultur', Berline, Germany, April 15-16, 2010.