22nd International Congress on Modelling and Simulation, Hobart, Tasmania, Australia, 3 to 8 December 2017
mssanz.org.au/modsim2017

Importance of the order of the modules in TransMob [Huynh et al., 2015]

M. Dumont [a], **J. Barthélemy** [b], **T. Carletti** [a] **and N. Huynh** [b]

[a] *naXys, University of Namur*
[b] *SMART, University of Wollongong*
Email: morgane.dumont@unamur.be

Abstract: Nowadays, a wide range of microsimulations are performed thanks to agent based simulations. This frame-work allows to model the interactions between agents and environment in a flexible way. Such models should be initialised with a given set of agents representing the initial population, and their interactions rules should be set. Very often the initial population of agents is obtained using a synthetic population generator.

TransMob is an agent-based model that aims to simulate transport and land use interdependencies for urban planning. TransMob contains two types of modules responsible for: the dynamics of the social structure within the population (ageing, divorces,...); and the travel behaviour of the individuals (assigning diaries, ...). In the current version of TransMob, the processes happening in the agents everyday life are modelled and performed in a specific order: first ageing, then death, divorce, marriage and finally giving birth.

This work focus on the impact of the ordering in which the different modules responsible for the update of the social structure are applied. This presentation aims at analysing the impact on the results if the order of the procedures is changed. For instance, how will the results change if the divorces are performed after the marriages? Let us denote the processes age, die, divorce, marriage and birth by 0, 1, 2, 3, 4 respectively. All possible orders are then given by the set of all permutations of the integers from 0 to 4. Thus, 120 different orders could be analysed.

However, if birth is applied before age, then in the first iteration, we will add new babies to the babies already in the initial population, resulting in an artificial peak of 1 year old agents in the first simulated year , 2 years old in the second simulated year, etc. For this reason, we only consider orders performing age before birth. This reduces the number of feasible orders to 60.

The analysis is performed using clustering and decision trees. The first results show that the place of ageing with respect to the place of death influences strongly the results. For instance, when ageing the agents is done before death, the final population is younger.

We aim to continue in this direction to further identify the consequences of choosing any particular order. The goal of this work and future research is to make scholars aware of the impact of a choice and provide them with findings to help them chose the best order for their application.

Keywords: *Spatial microsimulation, agent-based modelling, classification*

1 Introduction

Nowadays, a wide range of simulations are performed thanks to agent based modelling. This powerful framework allows to model the interactions between agents and environment in a flexible way. Each agent based model should be initialised with a given set of agents, the initial population, and their interactions rules. Very often (see for instance Ballas and Clarke [2001], Barthélemy [2014] and Cho et al. [2014]) the initial population is obtained using a microsimulation process.

The platform TransMob (Huynh et al. [2015]) is designed to simulate transport and land use interdependencies for urban planning and has been developed at SMART Infrastructure Facility of the University of Wollongong in the framework of agents bases modelling. This simulation contains two kind of modules : the dynamic of the social structure (ageing, divorces,...); and the assignment of each individual to the travel network (assigning diaries, ...).

This article will focus on the order in which the different modules responsible for the update of the social structure, are applied in the model. Dynamical processes of the everyday life can be modelled and performed in a specific order (first ageing, then death, divorce, marriage and finally giving birth). This article aims to analyse the impact on the results if the order of the procedures is changed. What is the impact on the results if we decide to perform the divorces before the marriages instead of the contrary? To reach this goal, the platform has been adapted to easily handle the reordering of the modules, hereby codified with integers, ”0, ..., 4”, using different orders. For example, if the input is ”0, 1, 2, 3, 4”, the considered order is age, die, divorce, marriage and birth.

All possible combinations of orders arrangements are then associated with the permutations of the numbers from 0 to 4. Thus, 120 different orders could be analysed. However, if birth is applied before age, then in the first year, we will add the new babies to the babies already in the initial population and it will make an artificial peak of 1 year old agents the first year, 2 years old in the second year etc. For this reason, we only consider orders performing age before birth. This reduces the number of admissible different orders to 60. We will use different analysis, a clustering and a decision tree.

This paper is structured in four parts. First, the algorithm using stochastic models, the stability of the process is checked in Section 2. For this purpose, the algorithm is executed several times with predefined random seeds. In a second phase, the influence of the order on the results is statistically tested in Section 3. Then, in Section 4, a k-means clustering is applied on the results of different orders to try to catch similarities in the final populations of several distinctly ordered simulations. The third part, Section 5 contains a linear correlation analysis to quantify the influence of the order and the random seed on the results. Finally, we expose some perspectives and conclude.

2 Stability

The stability of the algorithm with respect to the randomness introduced by the stochastic processes in action was checked in Huynh et al. [2015] for the permutation chosen in this article.

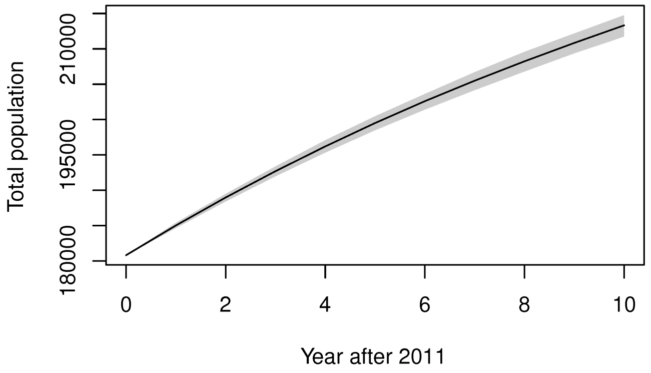

Figure 1. Stability - average and ranges of the simulated population. The solid line represents the average and the shaded zone the max-min.

The first step is to confirm the stability of the algorithm also once we introduce the modules with different orders. Figure 1 illustrates for each order and over 20 seeds, the average population size as a function of time since the beginning of the simulation, expressed in years, in black and the ranges in grey. We observe that all simulations lie very close to each other, which is a qualitative indication for the stability. The difference always increases with the number of simulated years, however let us observe that the values remains relatively small during the first ten years. The minimum, maximum and average is an important characteristics, but the information of the distribution between these two lines is also very useful. The number of men and women after 10 simulated years for 20 different seeds lies on Figure 2. The seed does not seem to influence these two indicators.

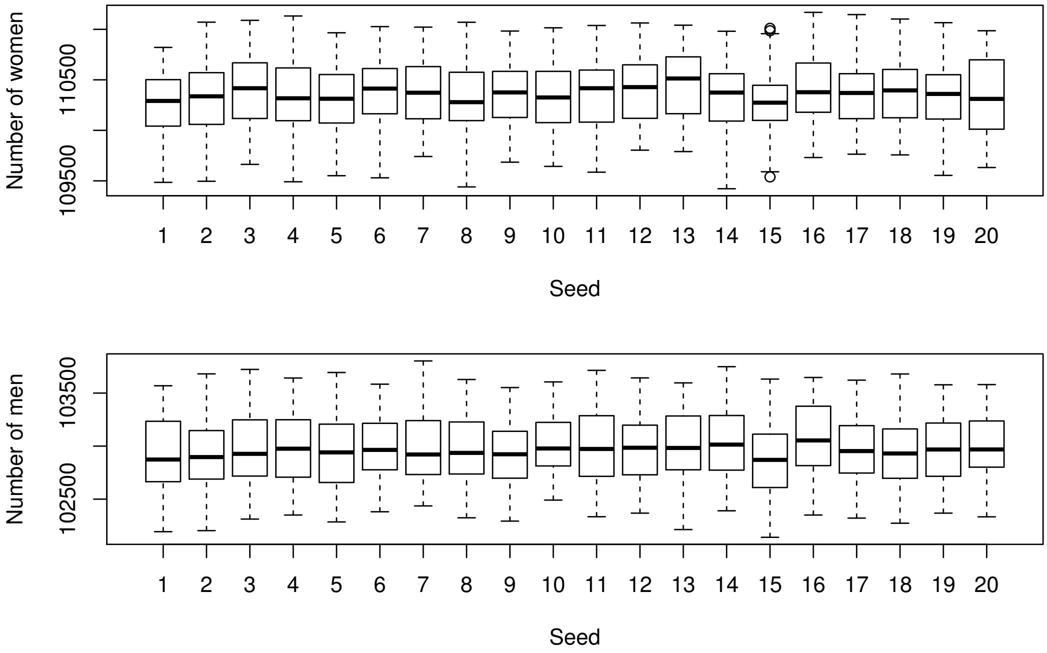

Figure 2. Stability per seed for each feasible order of the processes after ten simulated years

A Bartlett test confirms the homogeneity of the variances of the results through the seeds (with a p-value of 0.9868 for women and 0.9065 for men). Moreover, a Shapiro test indicates that for each seed, the distribution of the number of men and women follows a Gaussian law at level 0.01 (the smallest p-value,0.04, are obtained for seeds 13 and 20, all other seed being above 0.05). A statistical test with the null hypothesis ”The seed doesn’t influence the mean of the final population” has been executed implementing an ANalysis Of VAriance (ANOVA). The reference hypothesis is accepted at level 0.01 (p-value of 0.26 for women and 0.31 for men). Thus, the seed of the random number generator does not influence the final simulated population. We therefore conclude to the stability for the generations.

3 Influence of the order

To analyse the influence of the order on the predictions, the output of the algorithm has been redesigned to include the order of the processes, the seed (to check that they are not determinant of the classes), and the results in 2021 (10 years of simulations).

Before the designing of the influence of the order on the results, a statistical test is performed to ensure that the order significantly bias the final population. For this purpose, the homogeneity of variances of the size of the final population through each possible order is checked thanks to a Bartlett test. The p-value is 0.38 and confirms this homogeneity. The ANOVA obtains a p-value lower than 0.001. In conclusion, the number of simulated individuals after 10 years is different depending on the chosen order of the dynamical processes.

4 CLASSIFICATION

Knowing that different orders not necessary results in the same final population, we decided to compare each run by checking for the final year the total number of men, women and agents belonging to three age groups: less than 30 years old, between 31 and 60 years old, and more than 61 years old. A clustering is performed on these indicators and the final clusters are explained thanks to the order and/or the seed. To consider the orders, five variables are added to each simulation, the ”place” of each process. For example, if for a simulation we have in the order ageing, death, birth, marriage and finally divorce, the variables ”Place_age”, ”Place_death”, ”Place_birth”, ”Place_marriage” and ”Place_divorce” will be respectively 1, 2, 3, 4 and 5. If death is executed last, ”Place_death” becomes 5.

The number of classes, determined thanks to the elbow method using the k-means clustering, is two. We also tried in 3, 4 and 5 classes, but two seems to be the best number. The clustering is then performed to assign each combination to a class. Thanks to these classes, a decision tree can be constructed. In our case, the decision tree for two classes is illustrated in Figure 3.

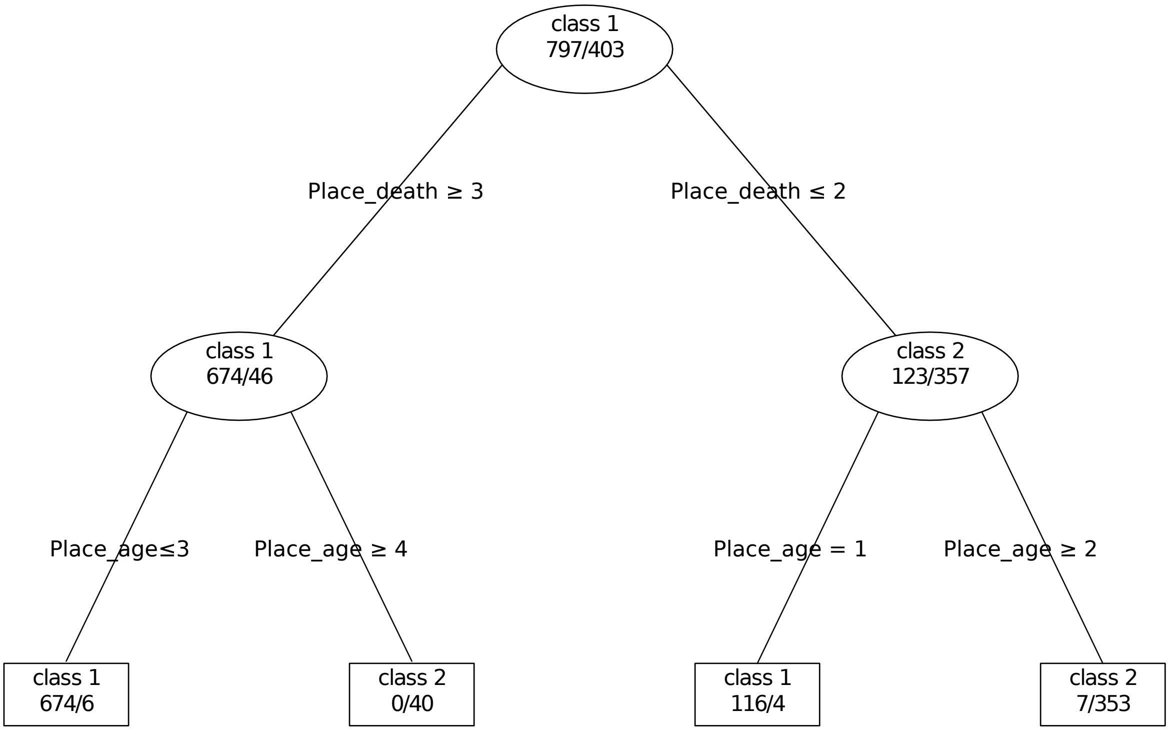

Figure 3. DecisionTree

The place at which the process implementing the death is the first thing to observe. If death is realised at the 3rd, 4th or 5th position, then we observe the place of ageing to chose the class. In each final class of the tree, we can see the number of simulations correctly assigned to the class. Indeed, we have in each cell the number of simulations assigned to each class by the classification algorithm and the class chosen by the tree. For example, for the first class on the left, 674 simulations were effectively assigned to class 1 by the k-means algorithm. However, 6 simulations do correspond to this place on the tree but were assigned to the second class. Simulations going into this branch are assigned to the correct class for $\frac{674}{674+6} = 99.12\%$ of the cases. Similarly, the other branches also give satisfactory results (in the order, good predictions in 100%, 96.67% and 98.05% of cases).

When we analyse more precisely each branch of this tree (from the left to the right) we obtain Table 1. We observe that in the first class ageing is always before death. The clustering shows that the seed was not determinant, only the order in which ageing and death occur is relevant.

Table 1. Decision rules

Class	Place of death	Place of ageing
Class 1	3, 4 or 5	1, 2 or 3
Class 2	3, 4 or 5	4 or 5
Class 1	2	1
Class 2	1 or 2	2, 3, 4 or 5

The impact of the order on the output is measured by analysing the results per class. Remember that as results, we record several indicators for the predicted population after 10 years of execution. Figure 4 shows that for both genders, the second class includes a larger population. Thus, when ageing is executed after death, the output of the algorithm is a larger population. This result is explained by the fact that when ageing, the probability to die becomes higher.

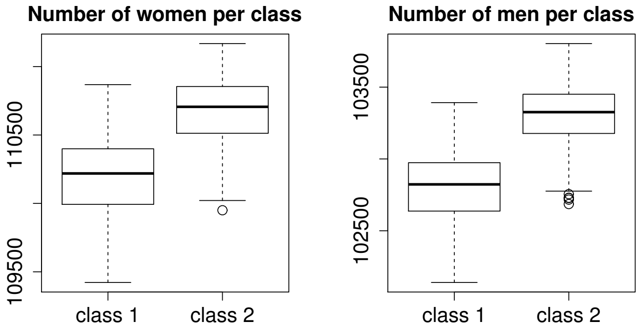

Figure 4. Boxplot of gender per class

The results for ages are illustrated in Figure 5. Both classes are similar in terms of number of individuals less than 30 years old. However, the older categories are more represented in the second class. So, when ageing is executed before death, the final population is younger. Intuitively, when ageing is performed before death, the rates of death (depending from age) becomes higher and the final population contains less elderlies. It should be noted that the k-means algorithm begins with a random classification. For this reason, we perform it 10 times to check the validity of the results. The 10 generations gave similar results (data not shown).

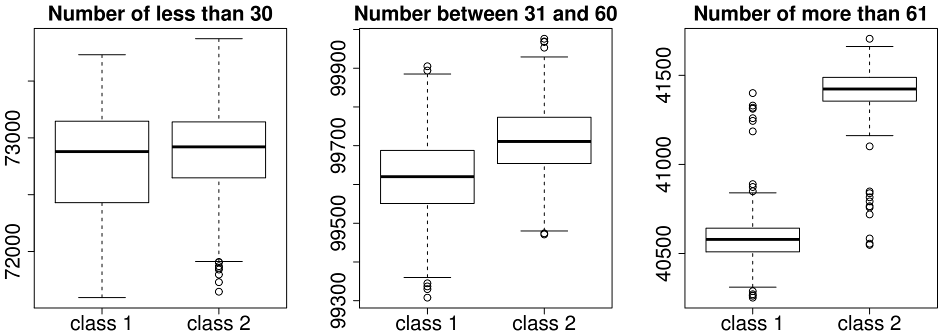

Figure 5. Boxplot of ages per class

Now that the tendency of the final population is explained, we'll focus on the evolution of these differences through the simulated years. For this purpose, the average age of the population is calculated per gender, per year, and for each simulation (20 seeds and 60 orders). Figure 6 indicates, per class of the order, the average of these average ages and the first and last quartiles. It confirms that the first class (having ageing before death) results in a younger population. The difference between the two classes begins early in the simulation and becomes higher from year to year. However, the curves don't really diverge and stay close to each other for these 10 simulated years. The ageing of the population is also observable on this graph, but the process tends to decrease the differences of average age between men and women in all simulations, independently of the order of the dynamical processes.

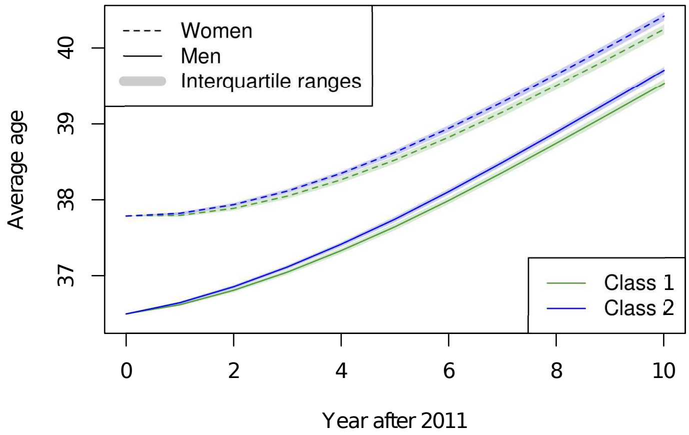

Figure 6. Average age of the population per gender for 20 seeds for each possible order (classified by the decision tree)

5 Correlation analysis

The k-means classification allows preliminary results, but it is sensitive to the scale (changing some variables from meters to kilometres for example could influence the results) of the data and give only convex classes. The following analysis of the Pearson correlation coefficients confirm these results. Table 2 contains the correlations between the results (after 10 simulated years) and the place of each process. To facilitate the reading of the coefficients, we coloured in red high correlations (more than 0.5 in absolute value), in green middle high correlations (between 0.3 and 0.5) and in blue the very low correlations (less than 0.05 in absolute value). A star means that the coefficient is not significantly different from 0.

Table 2. Correlations between the place of the process in the dynamical evolution and the results

	N_women	N_men	N_less30	N_31_60	N_more61
Place_age	0.51	0.55	-0.11	0.31	0.68
Place_death	-0.46	-0.52	0.01*	-0.33	-0.73
Place_div	0.16	0.14	0.23	-0.01*	0.01*
Place_marriage	-0.42	-0.36	-0.60	0.01*	0.00*
Place_birth	0.51	0.49	0.41	0.15	0.34

The place of age and death has an important role on the final population. Indeed, when ageing is performed later, the population of more than 61 years old will be higher (correlation of 0.678). On the contrary, when death arrives later, we simulate less elderlies 10 years after. We can see that the order has a higher influence on the number of men than women. The observations on this table confirm the results obtained by the clustering, but include also additional conclusions.

Marriages seem also to have a determinant role in the results. Indeed, if marriages arrive late, we have a smallest number of young individuals. The module of birth allows only married female to have a child. For this reason, if birth is before marriages, less female can have a baby and there are less young people.

Finally, the place of birth is also important. When birth arrives later, the population becomes bigger. It is fundamental to note that we forced birth to arrive after ageing. This implies that if birth is at the beginning of the process, ageing is also at an early place. We suspect that the high coefficients for the place of birth are biased by this decision.

6 Perspectives

We are working to a new way to avoid the ordering of procedure, say to chose if ageing is before death; even if relevant to the present work, we decided to postpone such result to a forthcoming publication. It consists in determining for each individual a birthday and considering the probability of each event as a linear combination of the probabilities of the age at beginning and end of the year. For example, if 2nd January the individual will become 25 years old, the probability to get married is:

$$P(Married) = \frac{2}{365} \times P(Marry|24yearsold) + \frac{363}{365} \times P(Marry|25yearsold)$$

This considers that the 2 first days of the years he is 24 and the other days of this years he is 25 years old.

On the same idea, death will be performed at the beginning and people dying this year will be assigned a day of death. For each other processes, if an event arrives this year to this person, a day for this event. This new event is validated only if death is after it.

7 Conclusions and Recommendations

In conclusion, different orders at which procedures are applied could be justified, for example, we could aim to perform divorces before marriages, to allow divorced people to get a new marriage in the same year. However, a married individual could also get divorced in the same year. The goal of our analysis is to make scholar aware of the impact of a choice and give ways to chose the better order.

The first results we present here shows that the order of the processes statistically influences the final population. The place of ageing with respect to the place of death influences the results. When ageing is before death (class 1), the final population is younger. We aim to continue in this direction to identify the consequences of each order and help future researchers to determine their optimal order.

Acknowledgments

- This research used resources of the ”Plateforme Technologique de Calcul Intensif (PTCI)” located at the University of Namur, Belgium, which is supported by the F.R.S.-FNRS under the convention No. 2.4520.11.
- SMART (especially Professor Pascal Perez and Tania Brown) that invited Dumont Morgane to start this jointly research.

References

Ballas, D. and G. P. Clarke (2001). Modelling the local impacts of national social policies: a spatial microsimulation approach. *Environment and Planning C: Government and Policy 19*(4), 587–606.

Barthélemy, J. (2014). *A parallelized micro-simulation platform for population and mobility behaviour - Application to Belgium.* Ph. D. thesis, University of Namur.

Cho, S., T. Bellemans, L. Creemers, L. Knapen, D. Janssens, and G. Wets (2014). Synthetic population techniques in activity-based research. In *Data Science and Simulation in Transportation Research*, pp. 48–70. IGI Global.

Huynh, N., P. Perez, M. Berryman, and J. Barthélemy (2015). Simulating transport and land use interdependencies for strategic urban planningan agent based modelling approach. *Systems 3*(4), 177–210.