Modeling of Hardy-Weinberg Equilibrium Using Dynamic Random Networks in an ABM Framework

Riccardo Tarantino[1], Greta Panunzi[2], and Valentino Romano[3](✉)

[1] Department of Humanities, University of Palermo, Palermo, Italy
[2] Department of Statistical Sciences, University of Rome "La Sapienza", Rome, Italy
[3] Department of Biological, Chemical, and Pharmaceutical Sciences and Technologies, University of Palermo, Palermo, Italy
valentino.romano@unipa.it

Abstract. Hardy-Weinberg equilibrium is the fundamental principle of population genetics. In this article, we present a new NetLogo model called "Hardy-Weinberg Basic model v 2.0", characterized by a strict adherence to the original assumptions made by Hardy and Weinberg in 1908. A particularly significant feature of this model is that the algorithm does not make use of the binomial expansion formula. Instead, we show that using a procedure based on dynamic random networks, diploid equilibrium can be achieved spontaneously by a population of agents reproducing sexually in a Mendelian fashion. The model can be used to conduct simulations with a wide range of initial population sizes and genotype distributions for a single biallelic autosomal locus. Moreover, we also show that without any mathematical formalism the algorithm is also able to confirm the prediction of Kimura's diffusion equations on the time required to fix a new neutral allele in a population, due to genetic drift alone.

Keywords: Hardy-Weinberg Equilibrium · Wright-Fisher model · Dynamic random networks · Agent-based model · NetLogo

1 Introduction

1.1 Historical Background and Conceptual Framework

Hardy-Weinberg Equilibrium is the first cornerstone of population genetics. The name of this principle is due to its two theorists, the English mathematician Godfrey Hardy and the German physician Wilhelm Weinberg, who independently modelled it and linked, using simple mathematical formalism, allele and genotype frequencies in an ideal population with certain characteristics [1, 2]. More specifically, Hardy's inspiration for showing the existence of the Equilibrium was a contemporary debate over a medical example (i.e., human brachydactyly). He considered a character linked to a single biallelic autosomal locus, fairly large population numbers, random mating, equal fertility for all individuals and even distribution of sexes among the three possible genotypes. Then, he derived

H. Cherifi et al. (Eds.): COMPLEX NETWORKS 2016 2022, SCI 1077, pp. 241–250, 2023.
https://doi.org/10.1007/978-3-031-21127-0_20

the expected genotype ratios at each generation based on allele frequencies. Wilhelm Weinberg published the binomial square principle the same year as Hardy but, in distinction to Hardy, who did no further work in population genetics, Weinberg went on to many other discoveries. Like Mendel's, his work went unrecognized for many years [3]. From the time of these achievements onwards, the importance of applying mathematics to genetic theory has been widely acknowledged, and it is still an active area of research [4, 5].

After these independent formulations, between 1928 and 1931 some other founders of populations genetics, namely Ronald Fisher, John Haldane and Sewall Wright, began to study the specific departures from the equilibrium, like those resulting by sampling distortions caused by limited population numbers [6]. The latter distortion is also known as random *genetic drift*, which was first investigated by the so-called Wright-Fisher model [7, 8]. In its most basic form, the Wright-Fisher model considers random drift as the only disruptive force of the HWE, due to the stochastic oscillations of allele frequencies resulting from random mating [9].

The simplest version of the Wright-Fisher model can easily simulate both HWE and drift, keeping in mind their sharing of almost all assumptions. Indeed, the model considers a randomly mating population consisting of a constant number of diploid hermaphroditic individuals which reproduce over discrete generations. New individuals are formed, at each generation, by random sampling with replacement of gametes produced by the parents, who die immediately after reproduction [10]. It is an example of a Markov process, where the future state $(t + 1)$ is only determined by its present state t (i.e., it is a memory-less process). In a biallelic model, there are two absorbing states, corresponding to the irreversible fixation or loss of one of the two alleles (i.e., when the frequency of one allele is 1 and that of the other one is 0) [11].

However, calculating the exact solution of the Wright-Fisher model without any simplification has been proved to be hard in practice. For this reason, starting from the 50s, the father of the Neutral Theory of Evolution Motoo Kimura developed a solution consisting in approximating the discrete WF process with a continuous time, continuous space diffusion process [12], which later brought him and Ohta [13] to develop the method of diffusion equations to estimate the *conditional* mean time of fixation for a single mutant neutral allele in a homogeneous allele pool ($4N_e$, with N_e = effective population size). The estimation of the time until fixation (expressed in number of generations) of a new allele can be achieved in different ways, and several studies have tried to confirm or improve the original Kimura's results with alternative approaches, such as those that make use of the coalescent theory [14, 15].

1.2 Agent-Based Modeling

NetLogo is a widely used software based on Agent-based Modeling (ABM) principles [16] and, during the past twenty years, many models have already been developed for processes and phenomena studied in a wide spectrum of disciplines ranging from Psychology to Biology and Chemistry. Such models are archived in a large Models Library and/or published in specialized journals. Nevertheless, few NetLogo models have attempted to properly implement HWE, and none of these models has fully implemented the assumptions made in the original mathematical formulation proposed by Hardy and Weinberg.

2 Materials and Methods

2.1 Implementation of HWE NetLogo Model

We have created the "Hardy-Weinberg Basic Model v. 2.0" using version 6.2.0 of NetLogo, which was downloaded from the NetLogo Home page of Northwestern University (https://ccl.northwestern.edu/netlogo/). Details on the implementation and features of this model can be found on the info and code tabs of the model and in the Results section of this article. For further information on the use of "Hardy-Weinberg Basic Model v. 2.0", the interested readership can contact the authors. The model has been uploaded on the NetLogo User Community. It can also be downloaded from the following link: https://www.dropbox.com/s/8sxvovb99bb04yg/HARDY-_1.NLO?dl=0.

2.2 Statistical Tests and Computer Simulations

To verify that the population generated in NetLogo is in Hardy-Weinberg Equilibrium, we used the Chi square goodness-of-fit test with continuity correction [17], X_c^2, for c = 0.5 [18]. The test is applied at each generation to compare the genotype frequencies obtained computationally and those expected, calculated on the basis of allele frequencies. The Chi squares for the sample are then:

$$X_c^2 = \sum_{i \geq 1} \frac{\left(\left|n_{ij} - E_{ij}\right| - c\right)^2}{E_{ij}} \text{ for c} = 0.5$$

The degree of freedom is the number of classes minus 1 and then minus the number of parameters estimated from the data, so with two alleles we consider only one degree of freedom. Hence, if we use a significance level of 5%, the tabulated, Chi square is 3.84. We reject the null hypothesis if the computed X_c^2 is greater than 3.84.

To establish if the average value of generations that we need to fix a newly appearing allele "*a*" is not different from the theoretical value ($4N_e$), we performed six different simulations with different population sizes (50, 60, 70, 80, 90, 100, 200, and 500 individuals/agents). These values of population sizes have been chosen to highlight the effects of genetic drift, whose action is stronger with small populations. On the other hand, genetic drift is acting with any finite population size. We did not try population sizes higher than 500 because this would take many more generations to reach fixation as well as requiring a longer computational time. Each simulation keeps track of the frequency evolution of the new allele (q) whose starting value is 1/2N (N = census number). All simulations were performed using the *BehaviorSpace* tool of NetLogo. Simulations stopped when $p = 0$ or $q = 0$, where p is the frequency of the more diffused allele "*A*". The results are stored in an Excel spreadsheet, where the number of generations required to fix or extinguish the "*a*" allele is reported for each iteration/replica. The cases of our interest are only the fixation events. We then compare by the one sample t-test the average value of generations calculated from a minimum of 20,000 replicates needed to fix the originally rare allele "*a*", with the theoretical value ($4N_e$) of the Kimura's model based on diffusion equations [13]. One-sample t-test is derived under the assumption that the population of data is normally distributed. Fortunately, even when data are not normally distributed, this method works well when the sample size is large enough. As we can observe in the following table, for each simulation we have a large sample size ($n > 100$).

Population size	50	60	70	80	90	100	200	500
Sample size (*n* of fixations)	168	193	116	105	130	107	58	42

With the one-sample t-test we compare, in each simulation, the average number of generations necessary to fix the allele "*a*" with the theoretical number $4N_e$ obtained with Kimura's diffusion approximation. Hence, for each simulation we have the following hypothesis test:

$$\begin{cases} \mu = 4Ne \\ \mu \neq 4Ne \end{cases}$$

where:

- $4N_e = 400$ if population size $= 50$
- $4N_e = 480$ if population size $= 60$
- $4N_e = 560$ if population size $= 70$
- $4N_e = 640$ if population size $= 80$
- $4N_e = 720$ if population size $= 90$
- $4N_e = 800$ if population size $= 100$
- $4N_e = 1600$ if population size $= 200$
- $4N_e = 4000$ if population size $= 500$.

Looking at the p-value, we reject the null hypothesis if $p \leq 0.05$ [19].

For each simulation we calculated the test statistic, the p-value and the confidence intervals considering the Student's t-distribution [20]. All results are computed using R 4.0 software [21]. We calculate the one sample t-test using the R software's function "t.test" [21].

3 Results

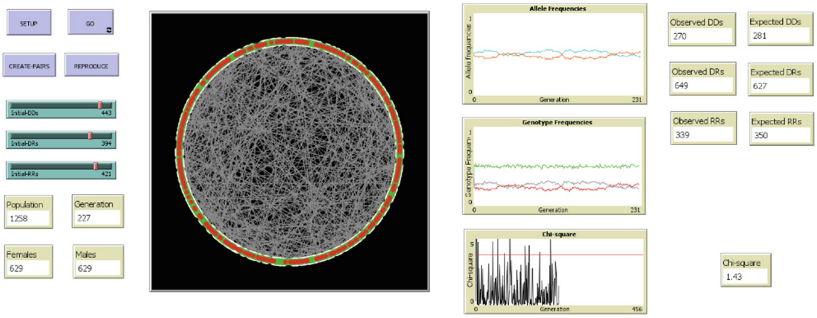
3.1 The Hardy-Weinberg Basic Model

The algorithm of the NetLogo HW model is presented in the code tab of the model. The agents in our model are diploid organisms with separated sexes. Each subset of agents is characterised by one of the three possible genotypes for a single autosomal locus (DD, DR, and RR). Two variables are considered for each agent: "sex" and "partner". Before running the model, the user may change the initial number of agents for each genotype, while the number of individuals for each sex is always kept constant with 1:1 ratio throughout simulations. In the setup phase, agents occupy fixed positions in the NetLogo world and are arranged in a circle. The random mating among all agents is implemented by links connecting, at each generation, two agents of opposite sex. In other words, the mating procedure was implemented by a dynamic random network whose topology continuously and synchronously change over time. The networks have many connected components with each component made of two nodes (N) corresponding to a mating pair. The process continues until no more unpaired agents are present (i.e., when the total number of edges is equal to the number of biunivocal pairs, N/2). Each mating pair generates two offspring, one male and one female. Gametes are not modelled, since reproduction is implemented by the expected genotypes reflecting Mendelian segregation ratios (e.g., there is a 50% probability to produce a DR offspring from a DR x RR cross). At the end of each cycle of mating and reproduction, both parents instantly die, that is, all the parental nodes are deleted and replaced by offspring nodes, that form a new random network. In the code, loop formation and elimination of random networks is implemented by executing the procedure “create-pairs” as follows:

```
to create-pairs
  ;; "singles" do not have any partner. Each single female creates
a link with a random single male. At the end of this procedure, all
females and males are paired
 let singles turtles with [ partner = nobody ]
 ask turtles with [ sex = "female" ] [
   if any? singles [
  if ( partner = nobody ) and ( any? other turtles with [ sex = "male"
] with [ partner = nobody ] ) [
     set partner one-of other turtles with [ sex = "male" ] with [
partner = nobody ]
     create-link-with partner
     ask partner [
     set partner myself
     ]
   ]
   ]
   ]
   tick
end
```

The choice of producing two sibs of opposite sex for each mating guarantees that sex ratio (1:1) and population size remain constant through generations. Thus, the population of agents that can be recreated in our model resembles a HW population because it is diploid, panmictic, large enough to maintain the equilibrium over time, has a 1:1 sex ratio, and has not overlapping generations.

A screenshot of the interface is shown in Fig. 1. Four buttons are used: (i) “setup” to initialize a simulation, (ii) “go” to run the simulation in a continuous mode, (iii) “create-pairs” to run it in a discrete mode, and (iv) “reproduce” to make couples generate their offspring. Three sliders are used to regulate the initial number of individuals for each of the three genotypes. Several monitors allow to keep track of (i) population size, (ii) number of agents for each sex, (iii) number of generations, (iv) observed frequencies for each genotype, (v) expected frequencies, and (vi) Chi square values. Expected genotype frequencies are estimated by the classic HWE formula $p^2 + 2pq + q^2 = 1$, where p and q are the frequencies of the "D" and "R" allele, respectively. p^2, $2pq$ and q^2 are, in this order, the expected number of "DD" (homozygotes), "DR" (heterozygotes), and "RR" (homozygotes). Two plots are used to display the evolution of allele and genotype frequencies. A third plot keeps track of change of the Chi-square values at each generation. In the latter plot, a red line is used to indicate the critical Chi square value of 3.84 for $\alpha = 0.05$ and df $= 1$.



Initial-DDs 443

Initial-DRs 394

Initial-RRs 421

Population 1258

Generation 227

Females 629

Males 629

Observed DDs 270

Expected DDs 281

Observed DRs 649

Expected DRs 627

Observed RRs 339

Expected RRs 350

Chi-square 1.43

Fig. 1. Interface of the NetLogo "Hardy-Weinberg basic model".

3.2 Time to Fixation of a New Neutral Allele

After calculating the mean time to fix the "*R*" allele and considering only simulations where the rare allele was ultimately fixed, we compare each of these average values with the theoretical value of $4N_e$. If variance in offspring number is greater than random expectation, Ne is smaller than N (i.e., the census number). On the contrary, if there is less random variation, N_e is greater than N. Finally, if all individuals have equal reproductive success, as in our case, then $N_e = 2N$ [10]. Thus, N_e values used for calculations are 2N. By applying the one-sample t-test for each N_e and sample size (i.e., Number of fixation events), we obtain the results shown in Table 1.

As in each simulation the p-value is greater than 0.05, we do not reject the null hypothesis (H_0: $\mu = 4Ne$) (see Methods). We can come to the same conclusion by observing the confidence intervals shown in Table 1 and Fig. 2; the theoretical value of $4N_e$ belongs to the confidence intervals at the 95% level. Table 1 and Fig. 2 also show that, as the effective population size increases, the width of the confidence intervals increases as well. This result is consistent with our expectation because, as the effective sample size increases, the sample size (i.e., number of fixations) decreases. In other words, the frequency of fixation events of the newly arisen neutral allele shows an inverse relationship with the total number of individuals in the population [14, 22].

Table 1. Statistical analysis of simulated fixation time observed in NetLogo for a new neutral allele.

DR/(DR + DD)	$4N_e$	t-test (df)	Average number of generations	95% confidence interval
1/50	400	−0.204 (167)	396.101	358.288 – 433.914
1/60	480	0.476 (192)	489.010	451.644 – 526.377
1/70	560	−0.198 (115)	554.129	495.439 – 612.820
1/80	640	0.398 (104)	655.086	579.914 – 730.258
1/90	720	1.554 (129)	773.069	705.488 – 840.651
1/100	800	−0.217 (106)	790.421	702.991 – 877.850
1/200	1600	0.805 (57)	1692.534	1462.277 – 1922.792
1/500	4000	0.617 (42)	4210.595	3521.06 – 4900.13

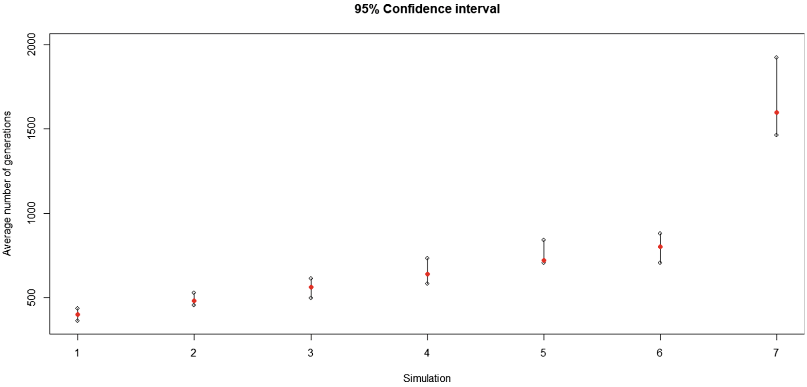


Fig. 2. The simulations are represented on the abscissa, while the number of generations are represented on the ordinate axis. Segments in (black) represent the confidence intervals at a 95% level. The red points represent the theoretical values $4N_e$, corresponding to N = 50, 60, 70, 80, 90, 100, 200. The point corresponding to N = 500 has not been included to improve C.I. visualization. For each simulation the theoretical values are included in the range, so we accept the null hypothesis in all cases.

Based on the above results, we accept the hypothesis that the outcomes of the simulations performed with our NetLogo model agree with the theoretical values estimated by Kimura and Ohta [13].

4 Discussion

In this study, our main aim was to create a NetLogo model of the Hardy-Weinberg Equilibrium using a complex system's perspective. Our approach was to develop an algorithm that does not make use of the binomial expansion formula of $(p + q)^2$. In this way, we show that—by using an ABM framework—diploid equilibrium can be achieved spontaneously as an emergent property of the complex system represented by the collective behaviour of agents reproducing in a Mendelian fashion. This feature distinguishes our model from other NetLogo models developed so far and also dealing with HWE (more about this below). Beside this peculiarity, the model is also strictly compliant to the implicit and explicit assumptions originally made in 1908 by the English mathematician Godfrey Harold Hardy and the German physician Wilhelm Weinberg. These assumptions can be summarized as follows:

- the organism is diploid; the considered gene is autosomal and bi-allelic (e.g., Brachydactyly);
- reproduction is sexual;
- generations are rigorously non-overlapping;
- sexes are evenly distributed;
- the allele and genotype frequencies are the same in males and females;
- mating is random (the population is panmictic);
- population size is large;
- migration is negligible, since gene flow by a population of different genetic structure would potentially change the allele frequencies in the original population;
- mutation can be ignored;
- natural selection does not affect the alleles under consideration.

Consistent with these premises, the simulation shows that equilibrium of genotype and allele frequencies is reached in one generation, and the population remains in equilibrium in successive generations, maintaining a constant sex ratio (1:1). Compliance to HWE can be verified by watching at the Chi square values displayed in the corresponding monitor and plot. Moreover, using the sliders, the number of individuals can be decreased up to a value where genetic drift is so strong to cause a rapid deviation from HWE, as can be verified in the Chi square plot when it repeatedly displays values above 3.84.

We shortly discuss here the recent "Hardy-Weinberg Equilibrium" model developed by Dabholkar and Wilensky [23], whose aims appear more similar to ours. These Authors aimed to model HWE in a population of rock pocket mice displaying alternative fur coat colors and state, in the Info tab, that their model performs according to the assumptions of HWE. However, by running their model, it can be easily verified that several agents survive for more than one tick/generation, that is, they have overlapping generations. This behavior makes that model non-compliant to the original HW model. In our model, we successfully captured the original level of abstraction of the HWE by using random dynamic networks. Therefore, a network approach can improve both adherence to the assumptions of population genetics' models and graphical representation of mating patterns in species with separate sexes, since all the current couples of nodes can be easily visualized by running the model in the discrete mode. The model does not rely on more specific network concepts, such as node activation or weight adaptation, since they are neither necessary nor relevant to analyze the collective behavior of our reproducing agents. However, it must be noted that the integration of ABM and network theory is presently a very active research area [24, 25].

Models dealing with finite populations must also deal with the effects of genetic drift. Thus, to explore further the performance of our model, we have used it to estimate the time of fixation for a new neutral allele because of genetic drift alone for different values of N_e. We have found that the results obtained with the simulations are consistent with those deriving from the approximated method of Kimura and Ohta [13], since the difference between the hypothesized value of $4N_e$ (with N_e = effective population size) and our calculated values is statistically not significant. Since the first formulation of this fundamental resolutive method of the Wright-Fisher model by Kimura and Otha [13], other independent approaches (e.g., coalescence theory [15]) have been used but, to our knowledge, ABM has never been used for this purpose. We are not reporting results for N values $>$ 100 because, above these values, simulations with NetLogo gradually become computationally intensive.

The extensive educational use of NetLogo today gives us hope that our model will be a useful tool for students undertaking college-level courses in population genetics. Using this model, teachers in this discipline have the opportunity not only to conduct their lessons with a multidisciplinary approach integrating genetics, networks, statistics, and agent-based modeling, but they can also encourage their students' systems thinking in studying and learning population genetics.

We plan to develop new versions of the model allowing to test for the effects of other evolutionary factors apart from genetic drift (i.e., selection, migration, mutation) in finite populations.

Acknowledgements. This work was funded by the Italian Ministry of Research, Teaching and University (MIUR) within the activities of the Prin project # 2020TACEZR_004 entitled: "Crossing the Sea: ancient and modern human genomes to study the evolutionary dynamics of Sicily, Sardinia and Corsica".

References

1. Hardy, G.H.: Mendelian proportions in a mixed population. Science **28**, 49–50 (1908)
2. Weinberg, W.: Uber den Nachweis der Vererbungbeim Menschen. Jahresh. Ver. Vaterl. Naturkd. Wurttemb **64**, 369–382 (1908)
3. Crow, J.F.: Population genetics history: a personal view. Annu. Rev. Genet. **21**, 1–22 (1987)
4. Sun, L., Gan, J., Jiang, L., Wu, R.: Recursive test of Hardy-Weinberg equilibrium in tetraploids. Trends Genet. **37**, 504–513 (2021)
5. Jagathesan, T.: Mathematical analyses in genetics and evolution. JDMS **09**, 138–142 (2022)
6. Mayo, O.: A century of Hardy-Weinberg equilibrium. Twin Res. Hum. Genet. **11**(3), 249–256 (2008)
7. Fisher, R.A.: On the dominance ratio. Bull. Math. Biol. **52**, 297–318 (1990)
8. Wright, S.: Evolution in Mendelian populations. Genetics **16**, 97–159 (1931)
9. Tataru, P., Simonsen, M., Bataillon, T., Hobolth, A.: Statistical Inference in the Wright-Fisher model using allele frequency data. Syst. Biol. **66**(1), e30–e46 (2017)
10. Charlesworth, B.: Effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. **10**, 195–205 (2009)
11. Hartl, D.L.: A Primer of Population Genetics and Genomics, p. 150. Oxford University Press, Oxford (2020)

12. Messer, P.: Neutral models of genetic drift and mutation. In: Kliman, R.M. (ed.) Encyclopedia of Evolutionary Biology, pp. 119–123. Academic Press, Oxford (2016)
13. Kimura, M., Ohta, T.: The average number of generations until fixation of a mutant gene in a finite population. Genetics **61**(3), 763–771 (1969)
14. Otto, S.P., Whitlock, M.C.: Fixation probabilities and times. In: Encyclopedia of Life Sciences. Wiley, Hoboken (2005)
15. Greenbaum, G.: Revisiting the time until fixation of a neutral mutant in a finite population—a coalescent theory approach. J. Theor. Biol. **7**(380), 98–102 (2015)
16. Wilensky, U.: NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL (1999)
17. Emigh, T.: A comparison of tests for Hardy-Weinberg equilibrium. Biometrics **36**(4), 627–642 (1980)
18. Yates, F.: Contingency table involving small numbers and the x2 test. J. R. Stat. (Suppl.) **1**(2), 217–235 (1934)
19. Thisted, R.A.: What is a P-value. Departments of Statistics and Health Studies. The University of Chicago, Chicago (1998)
20. Bulpitt, C.: Confidence intervals. Lancet **329**(8531), 494–497 (1987)
21. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018). http://www.R-project.org/
22. Kimura, M.: On the probability of fixation of mutant genes in a population. Genetics **47**(6), 713–719 (1962)
23. Dabholkar, S., Wilensky, U.: NetLogo Hardy-Weinberg equilibrium model. http://ccl.northwestern.edu/netlogo/models/HardyWeinbergEquilibrium. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL (2020)
24. Bouadjio-Boulic, A., Amblard, F., Gaudou, B.: Dynamic agent-based network generation. In: 9th International Conference on Agents and Artificial Intelligence (ICAART 2017), Feb 2017. pp. 599–606. Porto, Portugal (2017)
25. Anderson, T.M., Dragicevic, S.: Network-agent based model for simulating the dynamic spatial network structure of complex ecological systems. Ecol. Model. **389**, 19–32 (2018)