DR. DANILO G MUNIZ (Orcid ID : 0000-0002-7455-0801)

Article type : Research Article

Editor : Dr. Holger Schielzeth

A multinomial network method for the analysis of mate choice and assortative mating in spatially structured populations

Running title: Mate choice in spatially structured populations

Danilo G. Muniz[1,2,5], Eduardo S. A. Santos[2,3], Paulo R. Guimarães Jr[2], Shinichi Nakagawa[4] and Glauco Machado[2]

1. Programa de Pós-Graduação em Ecologia, Departamento de Ecologia, Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil.

2. LAGE do Departamento de Ecologia, Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil.

3. BECO do Departamento de Zoologia, Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil.

4. Evolution & Ecology Research Centre, and School of Biological, Earth and Environmental Sciences, University New South Wales, Australia.

5. e-mail: danilogmuniz@yahoo.com.br.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/2041-210X.12798

Abstract

1. Mate sampling, whereby individuals cannot access all potential mating partners in a population, is a ubiquitous yet poorly explored process. Ignoring mate sampling may underestimate female choice because the smaller the sample taken by individuals of the choosing sex, the weaker the correlation between sexually selected traits and the mating success among individuals of the chosen sex. A main factor promoting mate sampling is the spatial distribution of individuals. Thus, including distances in models of mate choice should improve estimates of mate choosiness. However, spatial distances between individuals are pairwise variables and cannot be readily included in the models commonly used to investigate mate choice.

2. We address this limitation by proposing a multinomial network (MN) model of mate choice, and comparing its performance with a previously published binomial GLMM. Both models allow the inclusion of pairwise predictors, accommodating spatial distances between individuals in analyses of mate choice. We evaluated the performance of these models in detecting directional and assortative mate choice using different simulated datasets: with and without spatial information, and with and without spatial autocorrelation of male and female traits. We also took samples of different sizes from the simulated datasets to evaluate the models' performance when data are incomplete.

3. Using both models, the exclusion of spatial information underestimated mate choice. Small sample sizes from the simulated populations led to underestimated directional mate choice, whereas assortative choice estimates were unbiased. Taking larger samples increased statistical power, and confidence interval coverage of both models. Spatial autocorrelation decreased the power of both models, but the MN model was less affected by it.

4. We conclude that including space in analyses of mate choice increases our ability to detect and accurately estimate mate choice using observational data. The MN model is a powerful and flexible tool that should be used in studies of mate choice in spatially structured populations. Moreover, the model can be used to investigate choice in other contexts, such as floral constancy by pollinators and host plant selection by phytophagous insects.

Key-words: assortativeness, information filtering, missing data, sexual selection, social networks, preference function, preference strength, spatial structure.

Introduction

Mate choice is an important process promoting sexual selection and influencing the evolution of mating systems (Andersson 1994). Females are usually the ones exerting mate choice, and the traits used in mate choice vary widely among taxa. For instance, males can be selected based on the conspicuousness of their coloration, size-related traits, acoustic features of their calls, chemical composition of pheromones, or multiple suites of traits (Andersson 1994, Candolin 2003). Irrespective of the trait involved in mate choice, females cannot evaluate all males in a population and must choose their mates from a limited sample (Janetos 1980). Female sampling establishes an information filtering process, which is any process that makes information about potential options unavailable to the choosing individuals (Mossa et al. 2002). Moreover, mate search can be costly and risky (Lane et al. 2010). Thus, time spent performing mate search is limited, and the average number of potential mates sampled by a female can be low (Roff & Fairbairn 2014). In natural populations, one of the main factors determining the males available for a female is the spatial distribution of males, because a female can only choose to mate with a male once she

encounters him (Byers et al. 2005).

Information filtering makes the detection of mate choice in observational studies more difficult: the smaller the sample of males taken by females, the weaker the correlation between a sexually selected male trait and a male's mating success (Benton & Evans 1998). This weakening of the correlation between the sexually selected trait and mating success occurs because a female may be restricted to a sample of suboptimal males in respect to the trait under choice, and yet she will mate with a male. Although information filtering is probably common in natural populations, it is regularly ignored in the statistical analyses of mate choice studies (but see Schlicht et al. 2015). The inclusion of spatial distances between males and females in analyses of mating data would account for a major source of information filtering. However, it is not trivial to include spatial distances between individuals as a predictor variable in statistical models because these distances are a pairwise variable.

Pairwise variables are difficult to include in regular statistical models because these variables only make sense when considering two specific individuals. Since the sampling unit used in analyses of mating data is typically the male, there is no way of including distances between each male and each individual female in the model. This is a problem not only when trying to include spatial distances in a model of mate choice, but also when testing hypotheses about assortative mating, because the similarity between two individuals is also a pairwise variable. Assortative mating is a common pattern in nature, but it can be generated by multiple processes (Jiang et al. 2013). It is often assumed that patterns of assortative mating are generated by individuals choosing mates similar to them (assortative mate choice). However, there is no standard method for detecting or estimating assortative mate choice. A correlation between male and female traits is commonly used as a measure of assortativeness (Jiang et al. 2013). Yet, this measure cannot distinguish between

assortativeness generated by mate choice and assortativeness generated by other processes.

Here we propose a multinomial network (MN) model of mate choice that can accommodate multiple pairwise and non-pairwise predictor variables. We compare the performance of this model with a binomial generalized linear mixed model (GLMM) proposed by Schlicht et al. (2015). To test the performance of these models under different preference criteria, we ran spatially explicit individual-based simulations in which females (i) had directional preference for large male traits, (ii) performed assortative mate choice, and (iii) performed random mating. We analysed the datasets using the MN model proposed here, and the binomial GLMM. To test the models under the confounding effect of spatial autocorrelation, we ran simulations with all preference criteria in scenarios of spatial autocorrelation of phenotypic traits. Finally, we took samples of different sizes from the simulated datasets to evaluate the models' performances with incomplete data.

Terminology

We define a *model of mate choice* as a probabilistic model that estimates the probability of an individual mating with another individual, and as *spatially explicit* any model that includes the spatial position of individuals in the estimation of copulation probabilities. The MN model, and the binomial GLMM fulfil these definitions.

Multinomial network model of mate choice

General description

The MN model is based on concepts of network theory, according to which a mating system is a sexual network in which individuals are nodes and copulations are links (McDonald et al. 2013, Muniz et al. 2015). The sampling units are the observed copulations, which are links

in the network. We assume that in each copulation one individual acts as choosing agent, whereas the other individual is available to be chosen. Although we refer to the choosing individuals as *females*, and to the individuals available to be chosen as *males*, sex roles can be defined according to the study system.

The model considers that a female i can mate with any of M available males, and estimates a probability for each male based on a set of T predictor variables. The probability P_{ij} of female i mating with male j is calculated as:

$$P_{ij} = \frac{exp\left(\sum_{h=1}^{T} W_h \cdot x_{h(i,j)}\right)}{\sum_{k=1}^{M}\left[exp\left(\sum_{h=1}^{T} W_h \cdot x_{h(i,k)}\right)\right]} \quad \text{(Eq. 1)},$$

where $x_{h(i,j)}$ represents the h_{th} predictor variable that can be either an attribute of male j (*e.g.*, body size, age) or a pairwise variable between male j and female i (*e.g.*, spatial distance, phenotypic similarity, genetic relatedness), thus the subscript (i,j). The values W_h are model parameters ("slopes") that estimate the influence (or weight, hence W) of each variable in determining copulation probabilities. Positive values of W_h indicate that x_h increases copulation probability, whereas negative values of W_h indicate that x_h decreases copulation probability. When $W_h = 0$, x_h does not influence copulation probability. We assume that the effects of predictor variables on P_{ij} are independent of each other, and we model P_{ij} as the exponential of the sum of weighted effects of each variable divided by the sum of the same exponential for all M males available for female j. This division ensures that, for each copulation, the sum of the probabilities of female i mating with all available males adds to one. Thus, the probability of a male j being chosen by a female i, P_{ij}, depends on the predictor variables x_h describing both male j and all males available for female i, and the model parameters W_h.

To fit the MN model, one must calculate a probability P for each observed copulation, so that the identity of the chosen male in each copulation is the multinomial response

variable of the model. Assuming that observed copulations are independent data points, and following the likelihood principle, we can consider these probabilities as likelihoods (Royall 1997). Once the probabilities *P* are calculated, the negative log-likelihood *L* of the model is:

$$L = -\sum_{l=1}^{N} log(P_l) \text{ (Eq. 2)},$$

where *N* is the number of observed copulations, and *l* is an index of copulations. The MN model can be fitted using maximum likelihood procedures, and we provide R codes and implementation examples in the Supporting Information (SI). Categorical variables can be included as predictors without any change to Eq. 1, but they must be binary. A categorical variable with more than two states must be included in the model as multiple binary variables (*i.e.*, dummy variables).

Model assumptions

The MN model has four key assumptions. First, copulations are independent data points, so that decisions by one female do not influence other females. Second, choice is performed solely by females, and the process is comparative, *i.e.*, females evaluate several males, and choose one based on the information gathered during the evaluation. Third, space has a monotonic effect on copulation probability, so that copulation probability will either increase or decrease continuously with the distance between individuals. Fourth, spatial locations of individuals are known with precision, but the model does not assume females possess perfect information about all males. On the contrary, by considering the effect of space, the model assumes that spatially distant males may be unavailable. The model *does not* assume any specific spatial distribution of individuals, and *does not* include the effect of past copulations (*i.e.*, no mate copying, and males can mate unlimitedly). In the following

section, we show how to use random factors to relax the assumption of independency between copulations by the same individuals, and in the SI (section 1) we relax the assumption of unlimited mating by males. Finally, in some populations, males may experience a post-copulatory timeout, whereas in others, mate-copying may increase copulation probabilities of recently mated males. Such effects can be included in the model by using the mating success of male *j* at a given time as a predictor of its mating probabilities at subsequent periods (SI, section 6).

Multiple measures and random factors

We can relax the assumption that copulations are independent data points by adding hierarchical structure to the model. Hierarchical models incorporate multiple measures, via random effects, which are usually varying intercepts or slopes (Gelman & Hill 2006). A random effect of male identity can be added to the model as an intercept G_j that varies for each male. Adding this random effect to Eq. 1 we get:

$$P_{ij} = \frac{exp\left(G_j + \sum_{h=1}^{T} W_h \cdot x_{h(i,j)}\right)}{\sum_{k=1}^{N}\left[exp\left(G_k + \sum_{h=1}^{T} W_h \cdot x_{h(i,k)}\right)\right]} \qquad \text{(Eq. 3)},$$

where G_j represents differences in the males' mating probabilities that could be influenced by morphological, behavioural, or social variables not included in any of the predictor variables $x_{h(i,j)}$. As each copulation is a sampling unit, the probability P_{ij} of a female *i* choosing a male *j* can only be estimated once female *i* has mated, which explains why differences in overall female mating propensity are not included in the model. Random effects of female identity must be added as varying coefficients (slopes) $W_{h(i)}$. For example, suppose we are modelling mate choice based on two pairwise traits, $x_{1(i,j)}$ and $x_{2(i,j)}$, so that model parameters are W_1 and W_2. Now, suppose we add a random effect G_j of male identity,

and a random effect of female identity as varying W_1 coefficients. The equation of this model is:

$$P_{ij} = \frac{exp\left(G_j + W_{1(i)} \cdot x_{1(i,j)} + W_2 \cdot x_{2(i,j)}\right)}{\sum_{k=1}^{M}\left[exp\left(G_k + W_{1(i)} \cdot x_{1(i,k)} + W_2 \cdot x_{2(i,k)}\right)\right]} \qquad \text{(Eq. 4).}$$

In Eq. 4, we can interpret each $W_{1(i)}$ value as an individual female's preference for trait $x_{1(i,j)}$, whereas we assume that all females are equally selective for trait $x_{2(i,j)}$. In the SI (sections 4-5) we show how to use Markov-Chain Monte-Carlo to fit the MN model with random effects.

Binomial mixed model of mate choice

Schlicht et al. (2015) proposed a model of mate choice that can include pairwise variables, this model is a GLMM in which each possible mating pair in the population is a sampling unit. The binary response variable is the occurrence of copulation, and because each possible mating pair is a sampling unit, pairwise variables can be included in the model along with non-pairwise ones. The identities of males and females are included in the model as random effects, so that interdependence between multiple observations of the same individuals is accounted for. The equation for the probability P_{ij} of a female i mating with a male j is:

$$P_{ij} = logit^{-1}\left(G_i + G_j + \sum_{h=1}^{T} W_h \cdot x_{h(i,j)}\right) \qquad \text{(Eq. 5),}$$

where the intercepts G_i and G_j represent random effects of female and male identity, respectively, T is the number of predictor variables in the model, W_h are the model coefficients, and $x_{h(i,j)}$ are the predictor variables. Differences in females' propensities to mate are represented by the varying intercept G_i.

Individual based simulations

To test the efficiency of the MN model and the binomial GLMM, we generated datasets using individual-based simulations. The simulations are spatially explicit and mimic a simplified mating system in which females choose males comparatively. Each simulation represents one independent mating season of an entire population with 1:1 sex ratio. The population comprises 400 individuals occupying a continuous two dimensional, squared landscape of 1 x 1 arbitrary units. At each simulation, individuals receive random x and y spatial coordinates from a uniform distribution between zero and one. Each male can mate unlimitedly, and each female copulates with a single male, chosen among the males available within a radius $r = 0.1$ from her spatial coordinate. This is a simplified way to model a population where individuals have fixed home ranges through the mating season. The $r = 0.1$ value was calculated so that, on average, females sample six males. If there is no male within r, females simply mate with the closest male. We considered rigid boundary conditions where individuals at the borders of the landscape had less space to sample mates, as if they occupied a habitat border. Still, the average number of sampled males in the simulations was always close to six.

We ran simulations with three mate choice criteria: directional, assortative, and random. In the directional mate choice simulations, males have a phenotypic trait z and females prefer males with large z values. The probability of a male j being chosen is proportional to z_j^{β}, where z_j is the male trait and β measures preference strength (Edward 2015). In the assortative mating simulations, males and females have phenotypic traits z, and females prefer males similar to themselves. The probability of a male j being chosen by female i is proportional to $exp\left(-\beta \cdot \frac{|z_i - z_j|}{z_i}\right)$, where z_j is the male trait, z_i is the female trait, and β measures preference strength. In both scenarios, we set β to 2 so that female choice is moderate. In the random mating simulations, females simply mate randomly with any male

within radius r. In all simulations, z follows a truncated normal distribution with mean z_{mean} = 4, standard deviation z_{sd} = 2, and minimum value z_{min} = 1. We adopted a truncation at z_{min} = 1 to avoid negative trait values, and to prevent males with z_j values close to the mean from having values orders of magnitude larger than other males.

Spatial autocorrelation can confound statistical tests (Dale & Fortin 2002), especially in the study of assortativeness, because it generates assortative patterns when there is no assortative mate choice. To evaluate the effect of spatial autocorrelation on the models' performance, we ran simulations in which male and female phenotypic traits were spatially autocorrelated for all three mate choice criteria. We generated spatial autocorrelation by determining the mean from which z-values were sampled based on the spatial location of individuals using the formula:

$$z_{mean} = 1 + \left(\frac{s_{io}}{s_o} \cdot 4\right) \qquad \text{(Eq. 6)},$$

where s_{io} is the Euclidean distance of individual i from the origin of the squared landscape (coordinates $x = 0$, $y = 0$), and s_o is the maximum distance to the origin. We considered a value of $z_{min} = z_{mean}/4$ and a $z_{sd} = 1$. We ran 50 simulations for each mate choice criterion, with and without spatial autocorrelation of traits, totalling 300 simulations in six scenarios. We implemented all simulations in R 3.2.2 (R Core Team 2015).

Analysis of the simulation data

We analysed data from each simulation using both the MN model and the binomial GLMM, and evaluated the effect of incomplete data by taking samples from the simulated data. In each sampling, we randomly took a proportion f of all females in the population, and considered that the males with which these females mated were measured. If the number of males was lower than that of females, we sampled additional males until the number of

males and females measured was the same. Thus, we ensured that in each sampling a proportion *f* of the population was measured. We considered 10 *f*-values, from 0.1 to 1 with increments of 0.1, and performed all analyses once for every simulated population, for each *f*-value. Given that unmated individuals were less likely to be sampled, our sampling procedure is analogous to a missing at random situation, whereby one variable (mated x unmated) influences the probability of missing data in another variable (Nakagawa 2015).

Directional mate choice

We analysed the data from the simulations without spatial autocorrelation of phenotypic traits with and without the use of spatial information, so that we could access the effect of ignoring spatial data. We only analysed data from the simulations with spatial autocorrelation of phenotypic traits including spatial data because the objective was to measure the influence of spatial autocorrelation on the models' performance. In the analyses with spatial data, we included two predictor variables in both models: spatial distance between individuals, and male trait z_j. In the analysis without spatial data, the only predictor was male trait z_j. We did not include random effects in these models.

In the MN model, we considered that the probability of a female mating with a male decays exponentially (which is different from the simulation, where all males within *r* are equally likely to be chosen). To preserve the scale of the mate choice parameter β, we used the logarithm property $z^B = exp(B \cdot log(z))$, and log-transformed the values of z_j for the analysis. Thus, we calculated the probabilities P_{ij} as:

$$P_{ij} = \frac{exp\left(-A \cdot s_{ij} + B \cdot log\,(z_j)\right)}{\sum_{k=1}^{M}[exp(-A \cdot s_{ik} + B \cdot log\,(z_k))]} \qquad \text{(Eq. 7)},$$

where s_{ij} is the spatial distance between female i and male j, z is the male trait, and M is the total number of available males. Parameter A measures the importance of spatial limitation on mate choice, whereas B measures preference strength for z_j (*i.e.*, it is an estimator of parameter β from the simulation). Log-transformation of traits is not obligatory in the MN model, but it has the advantage of making the choosiness parameter more intuitively interpretable. Thus, a value of 1 represents linear preference, whereas values > 1 represent super-linear preference. The disadvantage of this procedure is that the log–transformation does not accept standardized values.

Assortative mate choice

For the simulations with assortative mate choice, we performed the same analyses as in the simulations with directional choice. The difference was that, instead of the male trait z_j, the predictor variable related to individual traits was the relative difference u_{ij} in the trait between each mating pair. We calculated u_{ij} as:

$$u_{ij} = \frac{|z_i - z_j|}{z_i} \qquad \text{(Eq. 8)},$$

where z_i is the female trait and z_j is the male trait. Thus, in the MN model, we calculated the probability P_{ij} of a female i choosing male j as:

$$P_{ij} = \frac{exp(-A \cdot s_{ij} + B \cdot u_{ij})}{\sum_{k=1}^{n}[exp(-A \cdot s_{ik} + B \cdot u_{ik})]} \qquad \text{(Eq. 9)},$$

where s_{ij} is the spatial distance between female i and male j.

Detection of female choice and power

We adopted a significance level of 5%, thus we considered that a fitted model successfully detected female choice when the 95% confidence interval (CI) of the coefficient of female choice did not include zero. We measured statistical power as the proportion of model fits that were statistically significant.

Coefficient calibration and confidence interval coverage

The expected value for the coefficient B in the MN model is the value of parameter β from the simulations. To estimate the expected values for the binomial GLMM, we ran 100 calibration simulations for each mate choice criterion. These simulations used the same β parameter values presented above, with no spatial autocorrelation of traits, no information filtering (*i.e.*, females had access to all males), and large population size (1,000 individuals), thus creating ideal conditions for analysis. We analysed these simulations as described above, and used the mean coefficient values obtained as the expected coefficient values. We used these expected values to estimate the confidence interval coverage (CIC) of the models, calculated as the proportion of fitted models in which the 95% CI of the coefficient of choosiness included the expected value.

Coefficient standardization and prediction error

To make coefficients comparable among models, we standardized model coefficients using the following equation:

$$D_l = \frac{B_l - B_e}{B_e} \text{ (Eq. 10)},$$

where D_l is the standardized coefficient, B_l is an estimated coefficient, and B_e is the expected

value. We estimated predictive accuracy of the models using the metric of mean squared prediction error (MSPE, Wallach & Goffinet 1989). To produce comparable measures, we calculated MSPE based on standardized coefficients as follows:

$$MSPE = \frac{1}{N}\sum_{l=1}^{N}(D_l^2) \text{ (Eq. 11)},$$

where N is the total number of fitted models, and D_l is the standardized coefficient obtained from each fitted model.

Random mating

In the analysis of random mating simulations, we also took samples from the simulated data. We always included the spatial information in the analysis because our goal was to explore the models' performance in the absence of mate choice. We analysed the random mating simulations without spatial autocorrelation of traits in the same way we analysed the directional choice simulations: testing the hypothesis of preference for large values of z_j. Furthermore, we analysed the simulations with spatial autocorrelation of traits as we analysed the assortative mating simulations: testing the hypothesis of assortative mate choice. We did not standardize the coefficients from these analyses because the expected value was zero. We quantified false positive results as the proportion of cases in which the 95% CI of the models did not include zero.

Results

Directional mate choice

Incomplete data made both models consistently underestimate female choosiness (Figs. 1A, D). With complete data, both models were powerful in detecting female choice, showing CIC close to 95% and low prediction error (Fig. 2). With moderate amount of data (from 30-

70%) the MN model was slightly more powerful, but the binomial GLMM showed CIC closer to 95%. With only 10% of the data, the MN model showed high prediction error, but with more data it was more precise than the binomial GLMM (Fig. 2A). The exclusion of space from the analysis made both models underestimate choosiness (Figs. 1B, E). The inclusion of spatial autocorrelation of traits did not alter coefficient estimates (Figs. 1C, F), but decreased the power of the binomial GLMM and the CIC of both models (Fig. 2). Both the MN model and binomial GLMM always detected the spatial effect in the mating process.

Assortative mate choice

Coefficient estimates by the MN model were accurate (Fig. 3A) even with incomplete data, whereas the binomial GLMM slightly underestimated choosiness (Fig. 3D). The exclusion of spatial data from analyses promoted consistent underestimation of choosiness (Figs. 3B, E), but spatial autocorrelation did not have a strong effect on the estimates (Figs. 3C, F). With more than 30% of data, the MN model showed low prediction error, but with 10% of the data it showed high error (Figs. 4A, D). CIC was consistently below 95% (Figs. 4C, F), and with spatial autocorrelation of traits, CIC was lower. Both with and without spatial autocorrelation of traits, the MN model was more powerful. Both models always detected the spatial effect in the mating process.

Random mating

In the analyses with and without spatial autocorrelation of traits, the mean coefficient generated by both models was close to zero in all sample sizes (Fig. 5), with the exception of the smallest sample size, for which the MN model was biased (Figs. 5A,B). In the analysis of the simulations without spatial autocorrelation, the proportion of false positive results was always below 5% for the binomial GLMM, and was above 5% for the MN model only in the smallest sample size (Fig. 6A). In the simulations with spatial autocorrelation, the binomial

GLMM maintained the proportion of false positive below 5%, but the MN model showed slightly higher proportion of false positives for several sample sizes.

Discussion

We showed that the inclusion of spatial distance as a predictor in the MN model and binomial GLMM increased statistical power, and produced unbiased estimates of choosiness. In the directional choice analysis, choosiness was underestimated when space was ignored. Similarly, incomplete data also underestimated directional choosiness. Spatial autocorrelation of traits increased prediction error and decreased power and CIC, but did not bias estimates of directional choice. In the assortative mating analysis, incomplete data did not bias choosiness estimates, but increased prediction error and decreased power. Moreover, the binomial GLMM was slightly biased and less powerful than the MN model. Yet, both models showed CIC consistently below 95%, despite the high power. With incomplete data, spatial autocorrelation of traits decreased power and CIC of both models, but the MN model was less influenced. Finally, in the random mating analysis, both models produced accurate estimates of female choosiness, and the binomial GLMM produced a lower proportion of false positives. Given our results, both models become highly powerful and accurate with sample sizes ≥ 150 copulations, and it is ideal to sample 70% or more of the studied population when mate choice is moderate. In populations with stronger mate choice, smaller samples sizes may be sufficient.

Differences in the performance of both models may be partially attributed to their assumptions. Each model assumes different females' sampling behaviour and preference function. In the MN model, females are assumed to access information about several males, and then perform comparative mate choice (best-of-*n* process). The binomial GLMM, by using a logit function, implicitly assumes a probabilistic threshold process, in which at each

male-female encounter, a female decides whether she will mate or not without using information about other males. Considering this difference in assumptions, the binomial GLMM performed remarkably well in the analysis of simulated data, which were produced using the comparative process. The difference in assumptions between the models should guide researchers when choosing between them.

Both models were effective in detecting assortative mate choice. Assortativeness can be generated by multiple processes, but assortative mate choice is especially interesting because it is a component of many speciation models (Kondrashov & Shpak 1988; Martins et al. 2013). Thus, the MN model and the binomial GLMM are useful because they can separate spatial effects from assortative mate choice. These effects will not always be independent because individual traits can be correlated with location (Minias 2014), and site choice can be based on neighbours' traits (Doligez et al. 2002). However, the explicit separation of space and choosiness in the models allows for more detailed descriptions of assortativeness. Moreover, the coefficients produced by these models can be used as comparative estimates of assortative choice. Finally, the flexibility of both models allows the inclusion of assortative choice for a trait and directional choice for another, so that mate choice based on multiple traits can be studied even when the criterion of choice varies between traits.

In social networks, assortativity is often studied using a network-level descriptor, and by comparing the observed value of this descriptor with a theoretical benchmark provided by null models (Hu & Wang 2009). Null models, in turn, test if the observed value departs from expected by a theoretical benchmark, and/or estimate the amount of divergence by computing *z*-scores (Ulrich et al. 2009). The MN model improves the estimates provided by previous approaches in two ways. First, it allows the estimation of additive effects of each non-pairwise predictor variable in a direct way, considering the effects of statistical error. Whereas null model approaches do not allow one to compute the effects of two or more

predictor variables independently (Manly 2006). Second, the MN model allows us to compute pairwise predictor variables that are not available when using null model approaches in which most of the input parameters to compute interaction probabilities are estimated at the network level (Hu & Wang 2009).

Although the spatially explicit models of mate choice proved to be powerful and accurate, they require detailed spatial data. However, these data are usually already available. When studying nesting birds, for instance, nest locations are frequently recorded, and spatially explicit models can be used to investigate extra-pair copulations (Stewart et al. 2010). The same is valid for animals that live in burrows, such as some crickets for which there are detailed spatial data (Bretman et al. 2011). When individuals do not build nests or burrows, one can adopt home range centroids as individual coordinates, and use home range overlap estimates based on kernel functions to calculate the probability that a female will find specific males (Formica et al. 2010). For animals without stable home ranges or located via telemetry, one can use different spatial coordinates for each observation period when copulations are recorded (SI, section 6). Data gathering in these scenarios may be laborious, but the results obtained with spatially explicit models provide a more detailed portrayal of the mating system, including quantitative estimates of choosiness and spatial limitation involved in the mating process.

Here we presented the MN model and binomial GLMM as spatially explicit models of mate choice, but they can be seen as general models of choice or association that accept pairwise variables as predictors. Although both models were designed with the problem of spatial coordinates in mind, they can be used without spatial information, as flexible tools to investigate choice including pairwise variables. The binomial GLMM, particularly, does not necessarily presuppose a choosing agent, and can be used to investigate associations other than copulations, including the effect of multiple predictor variables on association

probabilities. Thus, the binomial GLMM is conceptually similar to p* models, which are general network models (Anderson et al. 1999). The MN model, in turn, presupposes that in each interaction there is a choosing agent interacting with a chosen subject. Thus, this model can be used to investigate choice in other contexts, two of which we explore below.

The first example is related to floral constancy in pollination, a behaviour exhibited by pollinators that restrict visits largely to a single floral type. According to this foraging behaviour, a pollinator that visits a flower from one type increases its chances of visiting a flower from the same type in subsequent visits (Amaya-Márquez 2009). However, travelling long distances between the same flower type may be costlier than visiting different flower types growing close together. Thus, the spatial distribution of different flower types probably influences flower choice by pollinators, which ultimately influence their degree of floral constancy (SI, section 7). The second example is host-plant selection by phytophagous insects. Females of several species select plants to lay eggs based on features such as plant species, size or conspicuousness, presence and extension of previous leaf damage, and presence of conspecific eggs and natural enemies (Bernays & Chapman 2007). If phytophagous insects have access to multiple plants, the spatial distance between potential hosts may influence the degree of selectiveness. By using the MN model, it is possible to evaluate the influence of the biotic and abiotic features on host-plant selection, considering the spatial distance between potential hosts.

We conclude that the inclusion of space in the analysis of mate choice increases our ability to detect, and accurately estimate mate choice using observational data, even with strong information filtering. Both the binomial GLMM and the MN model can include spatial distances between subjects as a predictor variable in statistical analyses. The performance of both methods is similar, but they have different assumptions that should be considered when selecting the most appropriate method based on the study system's

characteristics. Finally, we presented examples to show that the MN model is a flexible tool that can be used to investigate other types of choice.

Acknowledgements

We thank Diogo Melo, Paulo Enrique Peixoto, Gustavo Requena, Rafael Raimundo, Regina Macedo, Alexandre Courtiol, Jana Vamosi, and an anonymous reviewer for comments on the manuscript, Nat Lim for logistic support, and Benedikt Holtmann, Carlos Lara, and Yu-Hsun Hsu for suggestions. DGM, GM, and ESAS were funded by the São Paulo Research Foundation (2012/50229-1, 2012/20468-4, 2013/13632-5, 2015/10448-4). SN was funded by a Rutherford Discovery Fellowship (New Zealand) and a Future Fellowship (Australia).

Data accessibility

The code of the simulations used in this manuscript is available from github at https://github.com/danilogmuniz/informationFiltering/releases/tag/1.0 (DOI 10.5281/zenodo.8475).

Authors' contributions statement

DGM, PRG, and SN designed the probabilistic model. DGM and GM designed the simulations. DGM, SN and ESAS designed the analysis of simulated data. DGM wrote all simulation and analysis codes. All authors wrote the manuscript and the supporting information.

References

Amaya-Márquez, M. (2009). Floral constancy in bees: a revision of theories and a comparison with other pollinators. *Revista Colombiana de Entomología*, **35**, 206–216.

Anderson, C.J., Wasserman, S. & Crouch, B. (1999). A p* primer: logit models for social networks. *Social Networks*, **21**, 37–66.

Andersson, M. (1994). *Sexual Selection*. Princeton University Press, Princeton.

Benton, T.G. & Evans, M.R. (1998). Measuring mate choice using correlation: the effect of female sampling behaviour. *Behavioral Ecology and Sociobiology*, **44**, 91–98.

Bernays, E.A. & Chapman, R.F. (2007). *Host-Plant Selection by Phytophagous Insects*. Chapman and Hall, London.

Bretman, A., Rodríguez-Muñoz, R., Walling, C., Slate, J. & Tregenza, T. (2011). Fine-scale population structure, inbreeding risk and avoidance in a wild insect population. *Molecular Ecology*, **20**, 3045–3055.

Byers, J., Wiseman, P., Jones, L. & Roffe, T. (2005). A large cost of female mate sampling in pronghorn. *The American Naturalist*, **166**, 661–668.

Candolin, U. (2003). The use of multiple cues in mate choice. *Biological Reviews*, **78**, 575–595.

Dale, M. & Fortin, M. (2002). Spatial autocorrelation and statistical tests in ecology. *Ecoscience*, **9**, 162–167.

Doligez, B., Danchin, E. & Clobert, J. (2002). Public information and breeding habitat selection in a wild bird population. *Science*, **297**, 1168–1170.

Edward, D.A. (2015). The description of mate choice. *Behavioral Ecology*, **26**, 301–310.

Formica, V.A., Augat, M.E., Barnard, M.E., Butterfield, R.E., Wood, C.W. & Brodie, E.D. (2010). Using home range estimates to construct social networks for species with indirect behavioral interactions. *Behavioral Ecology and Sociobiology*, **64**, 1199–1208.

Gelman, A. & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Hu, H. & Wang, X. (2009). Disassortative mixing in online social networks. *Europhysics Letters*, **86**, 18003.

Janetos, A.C. (1980). Strategies of female mate choice: a theoretical analysis. *Behavioral Ecology and Sociobiology*, **12**, 107–112.

Jiang, Y., Bolnick, D.I. & Kirkpatrick, M. (2013). Assortative mating in animals. *American Naturalist*, **181**, E125–E138.

Kondrashov, A. & Shpak, M. (1998). On the origin of species by means of assortative mating. *Proceedings of the Royal Society B: Biological Sciences*, **265**, 2273–2298.

Lane, J.E., Boutin, S., Speakman, J.R. & Humphries, M.M. (2010). Energetic costs of male reproduction in a scramble competition mating system. *Journal of Animal Ecology*, **79**, 27–34.

Manly, B.F.J. (2006). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 3rd edn. Chapman & Hall, Chicago.

Martins, A.B., Aguiar, M.A.M. & Bar-yam, Y. (2013). Evolution and stability of ring species. *Proceedings of the National Academy of Sciences*, **110**, 5080–5084.

McDonald, G.C., James, R., Krause, J. & Pizzari, T. (2013). Sexual networks: measuring sexual selection in structured, polyandrous populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **368**, 20120356.

Minias, P. (2014). Evolution of within-colony distribution patterns of birds in response to habitat structure. *Behavioral Ecology and Sociobiology*, **68**, 851–859.

Mossa, S., Barthélémy, M., Eugene Stanley, H.E. & Amaral, L.A.N. (2002). Truncation of power law behavior in "scale-free" network models due to information filtering. *Physical Review Letters*, **88**, 138701.

Muniz, D.G., Guimarães Jr., P.R., Buzatto, B.A. & Machado, G. (2015). A sexual network approach to sperm competition in a species with alternative mating tactics. *Behavioral Ecology*, **26**, 121–129.

Nakagawa, S. (2015). Missing data: mechanisms, methods and messages. *Ecological Statistics: Contemporary Theory and Application* (eds G.A. Fox, S. Negrete-Yankelevish & V.J. Sosa), pp. 81–105. Oxford University Press, Oxford.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/.

Roff, D.A. & Fairbairn, D.J. (2014). The evolution of phenotypes and genetic parameters under preferential mating. *Ecology and Evolution*, **4**, 2759–2776.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London.

Schlicht, L., Valcu, M. & Kempenaers, B. (2015). Spatial patterns of extra-pair paternity: beyond paternity gains and losses. *Journal of Animal Ecology*, **84**, 518–531.

Stewart, S.L.M., Westneat, D.F. & Ritchison, G. (2010). Extra-pair paternity in eastern bluebirds: effects of manipulated density and natural patterns of breeding synchrony. *Behavioral Ecology and Sociobiology*, **64**, 463–473.

Ulrich, W., Almeida-Neto, M. & Gotelli, N.J. (2009). A consumer's guide to nestedness analysis. *Oikos*, **118**, 3–17.

Wallach, D. & Goffinet, B. (1989). Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecological Modeling*, **44**, 299–306.

List of supporting information files

1. *Supporting_Information.pdf:* contains a small amount of additional results (Section 1), and detailed tutorials and worked examples of how to fit the multinomial network model using both real and simulated data (Sections 2 – 7).

2. *flowers.txt*, *switches.txt*, *tcouples.txt*, *tfemales.txt*, *tmales.txt*, *ucouples.txt*, *ufemales.txt*, *umales.txt*, *wcouples.txt*, *wfemalesxy.txt*, *wmales.txt*, *wmalesxy.txt*: files containing simulated data used during the worked examples contained in the *Supporting_Information.pdf* file.

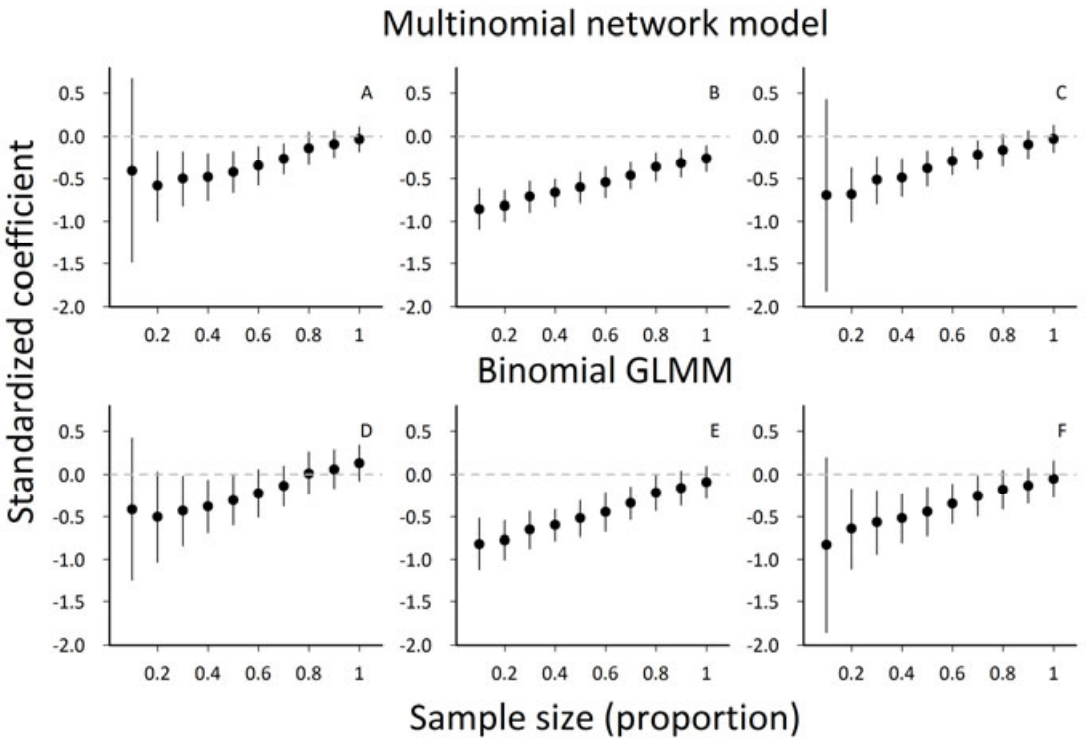

Fig. 1. Standardized coefficients (mean ± SD of 50 simulations) estimating the preference strength obtained from the analysis of simulated data in which females have directional preference for males with large traits. We show results obtained with samples of increasing size from a total population of 400 individuals. We analysed the data using two methods: (A-C) MN model, and (D-F) binomial GLMM. (A-B, D-E) Analysis of the simulations without, and (C, F) with spatial autocorrelation of phenotypic traits. Plots (B) and (E) show results when spatial information was excluded from the analyses. Grey horizontal lines indicate the expected values.

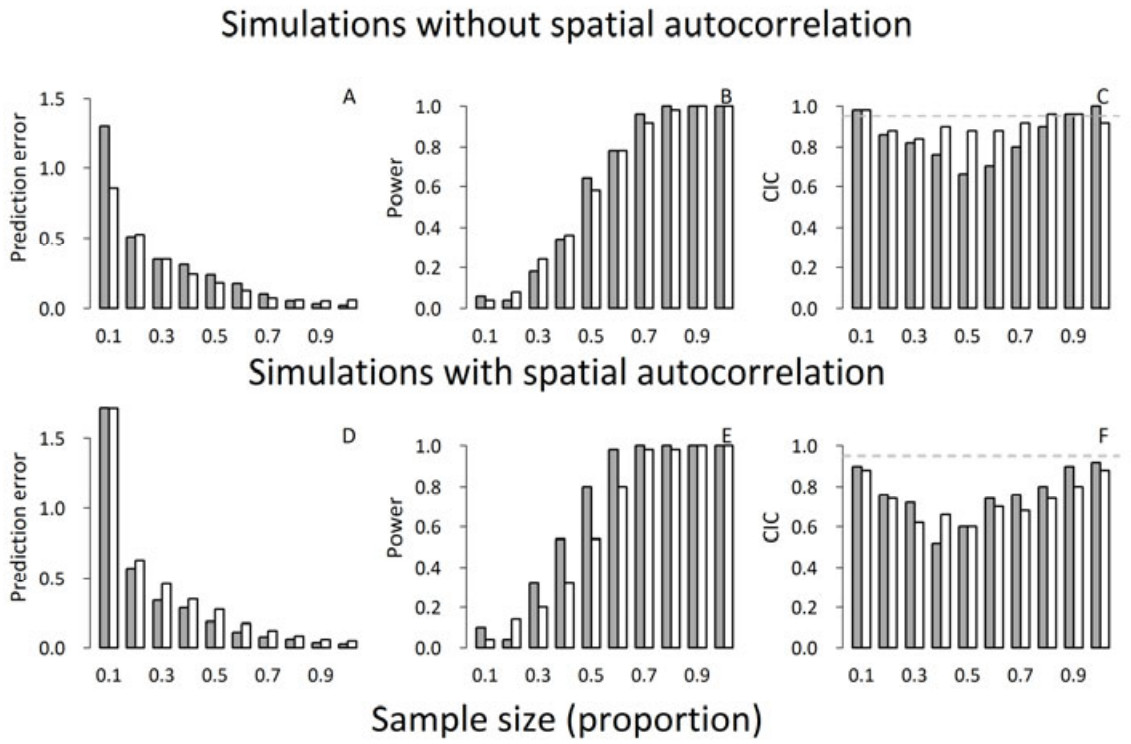

Fig. 2. Prediction error, statistical power, and confidence interval coverage (CIC) of the estimates of preference strength obtained from the analysis of simulated datasets in which females perform directional mate choice. We show results obtained by analysing samples of increasing size from a total population of 400 individuals. We analysed the data using two spatially explicit methods: MN model (grey bars) and binomial GLMM (white bars). (A, D) Prediction error measured as the standardized mean prediction error. (B, E) Power and (C, F) CIC of the models. The grey line in (C) and (F) highlights the 95% threshold.

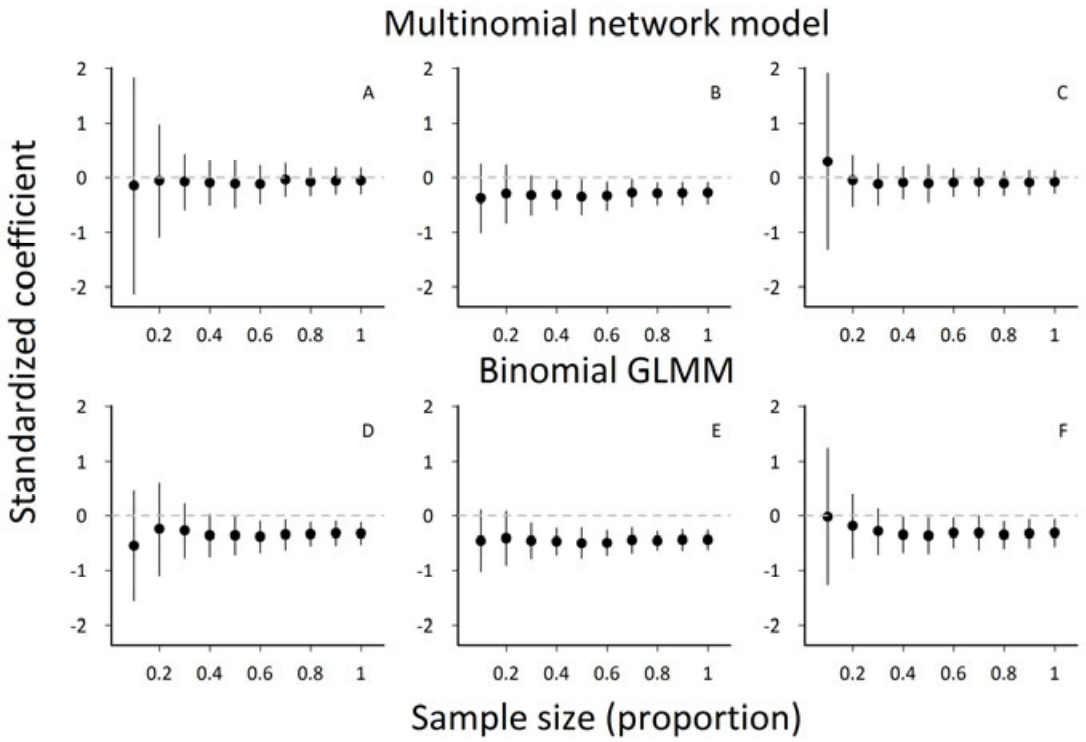

Fig. 3. Standardized coefficients (mean ± SD of 50 simulations) estimating the preference strength obtained from the analysis of simulated data in which females perform assortative mate choice. We show results obtained with samples of increasing size from a total population of 400 individuals. We analysed the data using two methods: (A-C) MN model and (D-F) binomial GLMM. (A-B, D-E) Analysis of the simulations without and (C, F) with spatial autocorrelation of phenotypic traits. Plots (B) and (E) show results when spatial information was excluded from the analyses. Grey horizontal lines indicate the expected values.

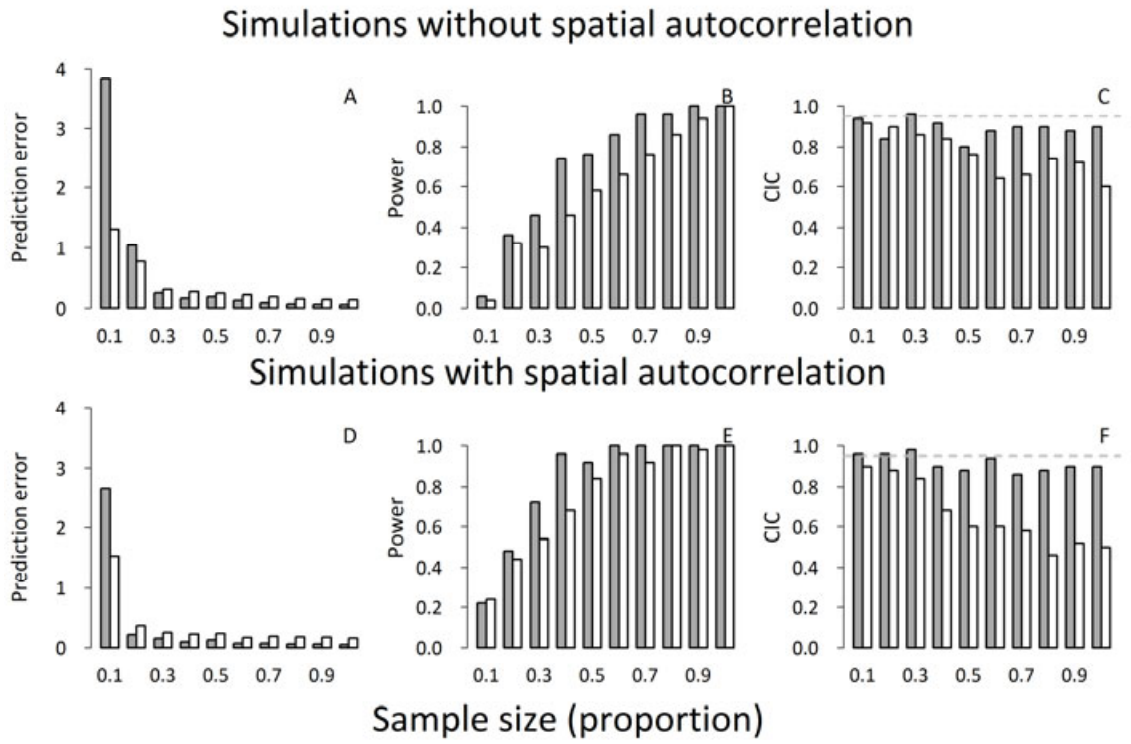

Fig 4. Prediction error, statistical power, and confidence interval coverage (CIC) of the estimates of preference strength obtained from the analysis of simulated datasets in which females perform assortative mate choice. We show results obtained with samples of increasing size from a total population of 400 individuals. We analysed the data using two spatially explicit methods: MN model (grey bars) and binomial GLMM (white bars). (A, D) Prediction error measured as the standardized mean prediction error. (B, E) Power and (C, F) CIC of the models. The grey line in (C) and (F) highlights the 95% threshold.

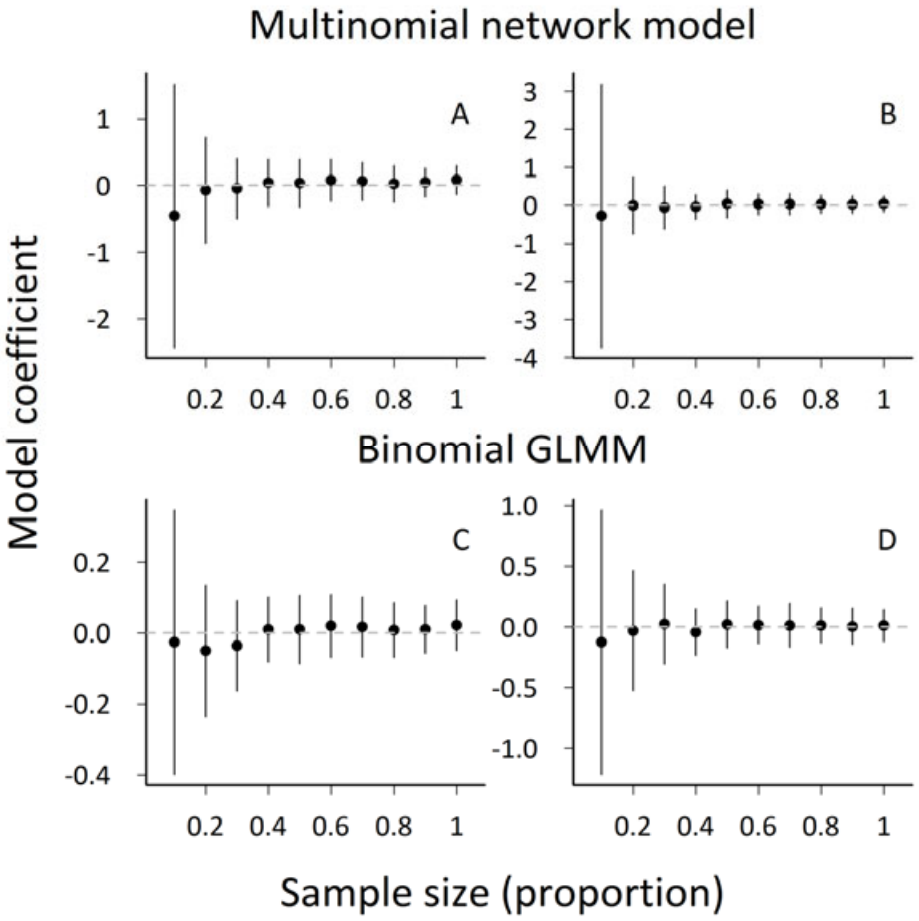

Fig. 5. Coefficients (mean ± SD of 50 simulations) estimating the preference strength from the analysis of simulated data in which females perform random mating. We show results obtained with samples of increasing size from a total population of 400 individuals. We analysed the data using two methods: (A-B) MN model and (C-D) binomial GLMM. (A, C) Simulations without and (B, D) with spatial autocorrelation of phenotypic traits to test the hypothesis of directional mate choice. Grey horizontal lines indicate the expected values.

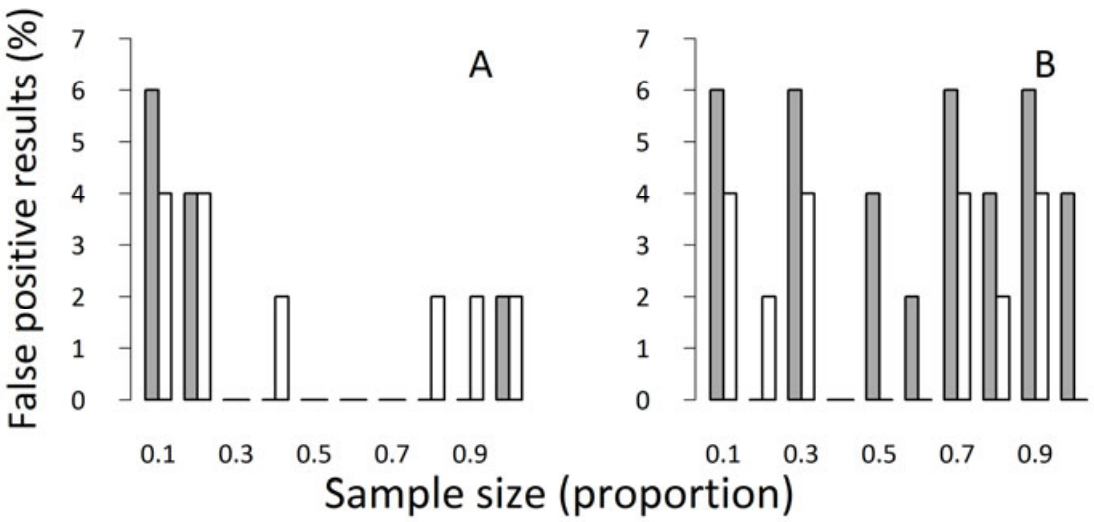

Fig. 6. Proportion of false positive results in the analysis of simulations with random mating in relation to the proportion of sampled individuals from a population of 400 individuals. Grey bars represent the MN model and white bars represent the binomial GLMM. (A) Simulations without and (B) with spatial autocorrelation of phenotypic traits to test the hypothesis of directional mate choice.